Writing Tests and Interpreting Test Statistics: A Practical Guide

I. Dwaine Eubanks
Lucy T. Eubanks

ACS DivCHED
Examinations Institute
Clemson University
Introduction

The purpose of this booklet is to help high school and college chemistry teachers write better classroom tests. The booklet grew out of several booklets that were used by ACS test committees in constructing high school and college tests for all fields of chemistry. Assessment is a vital part of teaching, and the ACS Examinations Institute has almost seventy years of leadership experience in providing the materials for meaningful assessment practices. It is our experience that chemistry teachers shy away from protracted theoretical discussions of assessment theory and seek practical and easy to apply approaches to constructing and analyzing their tests. We have tried to provide such a guide here. Individuals seeking a more esoteric, or comprehensive, discussion of test construction will wish to consult a standard textbook on educational measurement.

Principles of overall test design and item construction guidelines are presented in Section 1 of this booklet, starting on page 3. This section also contains guidelines for converting multiple-choice items to other commonly-used formats, such as true/false, short answer, and free response items. Section 2, starting on page 11, offers some suggestions for incorporating items with novel formats, constructing conceptual questions, and resources for laboratory practical questions. Overall test statistics and item statistics are powerful tools for improving the quality of chemistry examinations, and this subject is addressed in Section 3, starting on page 16. Definitions of commonly used statistical measures are presented, and guidelines for gathering and interpreting these measures are given.

These principles, suggestions and guidelines have proven helpful to the many chemistry teachers involved in writing examinations through the Examinations Institute of the American Chemical Society's Division of Chemical Education. We hope that you also find this information of genuine use in preparing and interpreting your classroom tests.

Your suggestions and comments about this publication are welcome. Please share your thoughts with us. You can reach us by mail, phone, FAX or Internet.

ACS DivCHED Examinations Institute
Clemson University
Clemson, SC 29631-1913
Telephone: (803) 656-1394 or 656-1249
FAX: (803) 656-1250
Internet: acsxm@clemson.edu

©1995 ACS DivCHED Examinations Institute. All rights reserved. Printed in USA.
Overall Test Design

A general format decision has to be made first. The most common formats are:
- multiple-choice; either four or five responses, with only one being correct
- short answer; where a word, phrase, or simple calculation suffices
- extended answer; requiring several sentences or solving multi-step problem
- practical items; involving gathering data or manipulating equipment

In addition to format decisions, the test maker must also decide the length of the test and the extent to which time stress is to be a factor in judging student performance. As the fraction of questions testing higher-order knowledge and abilities increases, the number of test items that can be used in a particular time frame decreases. The number of items and time allowed for most standardized tests has become well established over time, but teacher-made tests are often too long.

One useful way to begin to design a test is to construct a chart that lists concepts to be tested in one direction and level of knowledge to be demonstrated in another, as illustrated in the chart below. The entries are the number of items in each class (content objective and level of mastery).

### Sample Test Design Chart

<table>
<thead>
<tr>
<th>Objective</th>
<th>Understanding</th>
<th>Analysis</th>
<th>Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>TOTALS</td>
<td>15</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

Examinees should be given the opportunity to demonstrate various levels of competence in the subject matter being tested, not that they can simply regurgitate the information they have been fed. Each item being considered for the test should be analyzed to establish the level of mastery examinees must have to be successful in choosing the best answer, and important concepts should be tested at various levels of mastery. The American College Testing Program has developed one useful classification scheme, and test makers are well advised to analyze questions they write using similar criteria.

Understanding is the lowest level of knowledge tested. Examinees should be able to explain, describe, or identify the basic features of, and concepts related to, the provided information. At this level, examinees demonstrate that they are aware of basic facts, models, trends, classification schemes, principles, theories, and assumptions. They should also be able to recognize and display the information in graphs, figures, or diagrams.

Analysis is the next higher level of knowledge tested. At this level, examinees should be able to demonstrate that they can critically examine the relationships between information provided and the conclusions drawn or hypotheses developed. For example, they should be able to determine whether information or experimental results support, or are consistent with, a given hypothesis or conclusion. They should be able to base their conclusions on information presented in graphical or tabulated form.

Generalization is the highest level of knowledge tested. Examinees should be able to demonstrate the ability to generalize from given information to gain new information or broaden the context. Examples include development of conceptual or physical models, and prediction of outcomes when physical systems are manipulated. This level of knowledge is quite difficult to assess in multiple-choice format.

For most chemistry tests, the more questions that fall into the analysis or generalization categories the better, consistent with the goal of writing a test with which students can be reasonably successful. Once general decisions about the test have been made, the real work begins—that of constructing the requisite number of quality items distributed among the topic groups and levels of knowledge specified in the test design. Educational measurement specialists consider the issue of validity of test items (and of complete tests) to be central to the entire testing process. A valid test item should both address the desired scientific content (reflecting the test design criteria) and also be asked in a way that is appropriate for the instructor's objective. As an example, if we wish to determine whether a student really understands the concept of moles, we must construct items that can only be successfully completed if that concept is understood. The mass-mole calculation items, which are so common on chemistry tests, can be successfully solved using memorized algorithms. Successful students with these items often have a very imperfect understanding of the mole concept.

The merits of various item types (multiple-choice vs. free-response vs. short-answer vs. a practical task, for example) are often debated in content-free discussions, but each type lends itself best to certain content. There is little point in asking a free-response question about something that has a limited, closed set of alternatives, such as choosing the necessary pieces of laboratory equipment from a student locker to perform an acid-base titration. On the other hand, multiple-choice items are clearly inadequate if we wish to determine whether a student can successfully standardize a base. Here, there is really no substitute for doing the standardization in the laboratory. However, if the objective is to determine whether the student can do the required calculations, a short-answer question will reveal the most information—particularly if the student is also required to show his or her work.
Until recently, the selection of the "best" item type for particular content tended to be compromised by the pressures of making the items quick and easy to grade. Structured response items are, by far, the easiest and fastest to grade, and, in the United States, multiple-choice tests have become almost synonymous with any large-scale testing program. That situation is beginning to change, as test writers are becoming more adept at designing items that provide students with some freedom in structuring their own responses, while still maintaining machine scorability. Even lacking machine-scorability, unstructured responses can be fairly rapidly evaluated if they are very carefully constructed to enable the application of straightforward (and fast) scoring rubrics.

The main points of this section are to emphasize that, if the test is valid, every test item should both measure the intended attribute and be structured in an appropriate way. Unfortunately, there are no good, accessible measures of item or test validity. Validity simply cannot be measured quantitatively in most testing situations encountered by chemistry instructors. The only satisfactory way to assure content validity for a test is for it to have survived careful scrutiny by experts in the field, and instructors are always wise to invite colleagues to critique tests they write. Validity is discussed in greater detail in Section 2. At this point, we will consider constructing items for particular formats, beginning with multiple-choice items.

Constructing Multiple-Choice Items

Multiple-choice tests have several strengths that make this the format of choice for most objective examinations. Various levels of understanding—from simple recall to complex applications of principles—can be successfully tested (although actual multiple-choice items frequently fail to achieve this goal). Scoring multiple-choice tests is rapid and objective. Well-constructed tests are highly reliable, successfully measuring many student attributes the examiner may wish to test. Test statistics and item statistics are easy to calculate, making some measures of test quality quickly and easily done.

On the negative side, good multiple-choice tests are surprisingly difficult and time consuming to build. Every individual item must be well structured if the knowledge or ability in question is to be successfully measured. The test as a whole must be valid for all (or an appropriate sampling of) the knowledge or abilities expected by the examiner.

There are two major weaknesses of multiple-choice items for the test maker to consider. For one thing, students may be able to recognize a correct response that they would never have been able to devise. For another, students have no opportunity to demonstrate their thought processes in reaching their selected response. Test items can be framed in ways that minimize the negative aspects of both of these weaknesses.

Most test makers choose to write four-response items rather than five-response items. Generally, the extra time required to produce a fifth alternative of high quality for each question is not justified by the questionable benefit of having five responses.

For its own tests, the ACS Examinations Institute has necessarily adopted some policies regarding style that are arbitrary, but that do serve to maintain a consistent "look and feel." Alternative responses are labeled as (A), (B), (C), (D), and sometimes (E) for all items. Any test maker is well advised to use a consistent style with students, so that deciphering the test format does not become a factor in their performance. Units and abbreviations should conform to the recommendations of the most recent edition of The ACS Style Guide.

Syntax. Items may be based on either sentence-completion or response-to-question stems. Nearly all multiple-choice tests contain both formats. Responses to sentence-completion items should be followed by a period—except when the punctuation mark would be confused with a decimal point. No punctuation should follow each alternative of response-to-question items. Words in alternatives (including the first word) should never be capitalized unless they are capitalized in normal writing. These two common formats are illustrated in the examples that address common flaws.

Item Objective. A test item often is intended to address a single objective listed in the test design chart. Be careful to construct the item so that the intended knowledge or ability is really being tested. It is distressingly easy to end up with an item that appears acceptable at first glance, but actually focuses on an unimportant detail or an attribute unrelated to the objective of the item.

Consider an item that is intended to determine whether or not a student understands the notion of percent yield in a chemical calculation:

What is the percent yield if 0.50 mol of ethylene, C₂H₄, is reacted with 0.40 mol of H₂O to produce 0.35 mol of ethanol, C₂H₅OH?

\[
\text{C}_2\text{H}_4 + \text{H}_2\text{O} \rightarrow \text{C}_2\text{H}_5\text{OH}
\]

(\text{A}) 88\%  \hspace{1cm} (\text{B}) 80\%  \hspace{1cm} (\text{C}) 75\%  \hspace{1cm} (\text{D}) 70\%

As the question is written, it tests both understanding of percent yield and of limiting reagent calculations. The question would be better focused on the original objective if rewritten in this way:

What is the percent yield if 0.50 mol of ethylene, C₂H₄, is reacted with an excess of H₂O to produce 0.35 mol of ethanol, C₂H₅OH?

\[
\text{C}_2\text{H}_4 + \text{H}_2\text{O} \rightarrow \text{C}_2\text{H}_5\text{OH}
\]

(\text{A}) 88\%  \hspace{1cm} (\text{B}) 80\%  \hspace{1cm} (\text{C}) 75\%  \hspace{1cm} (\text{D}) 70\%

Language. Clarity of expression is a major factor in multiple-choice items. The sentence structure should be strong and grammatically correct. Students cannot resolve ambiguous language by reference to the larger context as they can in a textbook. Confusing language produces examinee responses that may not reflect their understanding of the content that the examiner intended to test.
It is important to avoid complex or awkward sentence construction. On the other hand, all necessary qualifiers to insure that the question has just one correct response must be included. The basic idea is to make sure that every word in the stem or responses is necessary and is the best choice. Avoid unnecessary words, nonessential specificity, and stereotyped phrases.

Place most of the wording in the stem, not in the responses. The examinee should be able to understand what is being asked after reading the stem, and should be able to formulate the rest of the sentence or the response to the question without looking at the proposed responses.

Common elements should be included in the stem rather than in the responses. If this appears difficult with a completion statement, switch to a question format (and vice versa).

Whenever possible, items should be stated in a positive form. Negative statements are often misinterpreted, leading examinees to choose unintended responses. If it is necessary to include a negative, emphasize it with italics, boldface, or capital letters. Never include negatives in both the stem and any response.

Relevance. Normally only relevant material should be included in the item. Irrelevant or extraneous material may cause a test item to be more difficult without any gain in assessing the examinee's understanding of the material. An exception to this practice occurs when the examiner wishes to ascertain whether a student can identify relevant material from larger compilations.

Verbal clues. Some grammatical constructions and words allow test-wise examinees to eliminate poor responses or to identify correct ones with little knowledge of the subject matter. Occurrence of a key word in both the stem and one of the alternatives, for example, is a surprisingly common clue. Most of the problems relating to verbal clues occur in the relation of responses to the stem.

Item independence. Every item should be independent of every other item of the test. Information given in one question often provides a clue to the correct response in another item. This does not preclude basing several items on the same data or scenario.

Responses. Every alternative should be plausible and should anticipate likely misconceptions or computational errors. The responses should be arranged in a sequence that gives no clue to the correct answer. Generally, this means sequencing numeric responses or single letters in ascending or descending order. The alternatives should be parallel in grammatical form because non-parallel responses often reveal the correct answer or obviously wrong answers.

Everyone who writes tests tends to place the correct answer in one position in the order of answers more frequently than in the other positions, so the response order must be deliberately randomized to prevent test-wise students from scoring above their actual knowledge level.

Test makers usually spend more time and intellectual effort crafting the correct response than they do in developing the distractors. Consequently, the correct response often stands out. Test-wise examinees, for example will key on a much longer (or much shorter) response as the most likely answer. Either the response is made longer because extensive qualification is required to assure that the response is true, or the response is shorter because extensive qualification is required to assure that the distractors are indeed false.

If any language from the stem is repeated in an alternative, it is more likely to be in the correct response. On the other hand, vague references to nouns in the question stem generally signal false statements.

Always and never are most often used in false statements, and test-wise examinees key on these words. By the same token, "All of the above" and "None of the above" allow examinees to select a response based on the weakest remaining alternative.

Vocabulary often allows an examinee to choose a correct response or eliminate incorrect ones. If, for example, technical vocabulary that has not been introduced in a course appears in an alternative, it is always false. Likewise, alternative synonyms in parallel constructions must be incorrect responses.

When opposite alternatives appear, the correct response must be one of the two if they have any relationship to the stem, and the examinee can ignore the other alternatives. A similar problem exists with alternatives that overlap (<10, <20, <50, <100, for example), and the overlap limits the choices from which the correct answer is chosen. If a particular response (say, <20) were correct, those that overlap it (<50 and <100) would also have to be correct; hence, the only unoverlapped response (<100) must be correct.

This discussion can only serve to make chemistry educators who construct multiple-choice tests aware of some of the problems and opportunities associated with constructing good tests. Each of the potential difficulties could be illustrated with many examples of what to do and what not to do. Anyone examining tests they have written (or ACS DivCHED examinations, for that matter) will find many examples of these "rules" being violated. We must accept the fact that every test we write could be better, and that test construction is an iterative process in which perfection is never achieved.

Again, please remember that good items do not necessarily produce a valid or reliable test, but it is also true that good tests are not built from poor items. Accomplished test makers address both the overall design of the test and the quality of every item that goes into the test.

Constructing Short-Answer Items

Many times you will wish to use questions that are in other than four- or five-response multiple-choice format. Instructors will wish to either modify the existing questions or write entirely new items for those situations. Many of the design considerations relating to the stem of multiple-choice tests apply equally to other formats, and should be kept in mind when producing new items.
good multiple-choice items are already available, it is often easier to modify them to produce the desired format than it is to start afresh.

As an example of converting a multiple-choice item to other formats, consider this question, which is readily re-cast into true/false, short-answer, and free-response formats.

When metallic potassium is placed in water,
(A) the solution becomes mildly acidic.
(B) the potassium readily dissolves without reacting.
(C) a yellow, crystalline precipitate is observed.
(D) a gas is evolved.✓

True/False Format

Many multiple-choice items are, in reality, four or five true/false questions rolled into one. Even when they are not, it is often possible to obtain several true/false questions from a single multiple-choice question. Here, for example, a true/false question derived from the previous question might read as shown; additional possibilities are presented by the other responses.

True or False: When metallic potassium is placed in water, the solution becomes mildly acidic.

This is false, of course, but one has to be careful with converting multiple-choice questions into true/false questions. There are usually three false questions for every true one created from a four-response multiple-choice question!

Short-Answer Format

Many times, responses can simply be omitted and an acceptable short answer question results. This is not the case here, however, for the stem by itself does not offer enough guidance to elicit a limited range of brief answers. At its best, an item—even though tagged as short-answer—will channel the student’s thinking and provide a clear opportunity to demonstrate relevant knowledge.

The conversion of a multiple-choice item into a short-answer item usually makes the item more difficult for students. Often, the question may be rendered easier or more difficult with slight changes in wording. The level of difficulty can also be influenced by the use of symbols, diagrams, and graphs. Graphical elements in an item may make it easier for some students and harder for others, depending on their learning styles.

Short Answer: Write the formulas for the two products that are formed when metallic potassium is placed in water.

Free-Response Format

Free response items are judged by many instructors as being the only viable question type for most situations, but they can be as difficult to write well as they are to grade fairly. The important, and daunting, task is to anticipate the range and levels of student responses. The question in the example under discussion can easily be used as the basis for an extended essay question if the reasons for the reaction between potassium and water were asked for. The expected criteria for the extent of the response should be specified by the teacher so the student is not left guessing as to the required level of detail expected in the response to the question. Here is one suggestion for an item that is readily scored while allowing the student some latitude for response.

Free-Response: Using a net ionic equation and specifying all physical states, explain what happens when metallic potassium reacts with water.

Section 2. Alternative Question Formats and Styles

For many teachers, the holy grail for chemistry assessment would be test items that could be objectively scored without the limitations and shortcomings of multiple-choice items. Nowadays, it is feasible to design machine-scorable, structured-response items that place less emphasis on the application of algorithms and a greater emphasis on higher-level reasoning skills. This has become possible with the widespread replacement of stand-alone optical scanners by OMR data terminals interfaced to microcomputers. Technological developments in machine scoring afford new opportunities and can decrease reliance on single-answer questions, which may reflect a student’s ability to choose correctly rather than basic understanding of chemical principles and phenomena. This section addresses some of the options that chemical educators have to widen the format and styles of assessment, while retaining practical approaches that a teacher can implement. If these alternatives are unfamiliar to you, do not despair at the time required for your first efforts to write new types of assessment items. Ideally you will have a time to try these in a workshop situation with your peers; group work can be very beneficial for teachers as well as for students. Also remember that students will need to become familiar with these formats and styles in a non-threatening situation before using them in a “real test”!

Grid Question Format

A frequent criticism of paper-and-pencil tests is that many items are more a test of reading than they are an opportunity for a student to demonstrate understanding. Grid questions can lessen the dependence on the written word by presenting essential information in an array that contains graphs, formulas, words, pictures, or diagrams—or it may overlay a single large graphic, like looking through a window with panes. The grid serves as a focusing device, making the search through the choices more efficient. The number of elements in the grid, and the complexity of those elements, should be appropriate for the
age and experience of the students. Here are two examples, each of which assumes that students know the ground rules (here to choose all correct answers).

<table>
<thead>
<tr>
<th>Which monomer(s) can be used in the process of addition polymerization?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) [ \begin{array}{c} \text{H} \ \text{C} \ \text{H} \end{array} \quad \text{CH}_3</td>
</tr>
<tr>
<td>(B) [ \begin{array}{c} \text{C} \ \text{C} \end{array} \quad \text{Cl}</td>
</tr>
<tr>
<td>(D) [ \begin{array}{c} \text{H} \ \text{H} \end{array} \quad \text{H} \quad \text{NH}_2</td>
</tr>
</tbody>
</table>
| (E) \[ \begin{array}{c} \text{F} \\ \text{C} \quad \text{C} \quad \text{F} \end{array} | (F) \[ \begin{array}{c} \text{H} \\ \text{C} \\ \text{C} \quad \text{H} \end{array} 

This grid question illustrates several important points about design. In this case, where there are four correct answers (B, C, E, and F), scoring necessarily involves partial credit as well as penalty scoring. The most popular scheme for this example would be to award 1/4 credit for each correct answer and to deduct 1/2 credit for each incorrect one. (If a student chose all six responses, these credits and debits would sum to zero, just as if the student had marked no responses.) Most instructors elect to award zero in preference to a net negative score for an item. (Here, for example, A, B, and D would sum to −3/4.)

<table>
<thead>
<tr>
<th>Which graph(s) represent the relationship between volume and pressure for a fixed sample of an ideal gas being held at constant temperature?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) [ V \quad \downarrow \quad P \rightarrow</td>
</tr>
<tr>
<td>(B) [ V \quad \downarrow \quad P \rightarrow</td>
</tr>
<tr>
<td>(C) [ V \quad \downarrow \quad P \rightarrow</td>
</tr>
<tr>
<td>(D) [ V \quad \downarrow \quad 1/P \rightarrow</td>
</tr>
<tr>
<td>(E) [ V \quad \downarrow \quad 1/P \rightarrow</td>
</tr>
<tr>
<td>(F) [ V \quad \downarrow \quad 1/P \rightarrow</td>
</tr>
</tbody>
</table>

This grid makes effective use of graphs as the grid elements. As with the previous example, there is a chance for each student to demonstrate at least partial knowledge of the relationship, but not all of them will recognize that, in addition to choice (B), choice (F) also represents the inverse proportionality of pressure and volume.

Only one question was asked here based on each of the grids, but grid questions can often be used as the basis for several questions. To fully utilize the power of the grid, most of the questions asked should be designed to have more than one correct (or partially correct) answer. This allows students to demonstrate partial knowledge.

Examples of using grids for laboratory assessment include such things as (1) a grid overlay of a laboratory set-up, with the student expected to identify the grid within which a specific thing happens (such as the location in a distillation set-up where the liquid and vapor are in equilibrium) and (2) Individual pieces of equipment in the grid cells with the student expected to select pieces to join for a particular set-up. Here is an example of the second type:

<table>
<thead>
<tr>
<th>Each of these pieces of equipment is necessary for fractional distillation. Identify all the other pieces (by letter) that are connected to each piece.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) [</td>
</tr>
</tbody>
</table>

The responses (A: E, F); (B: G); (C: D, E, G); (D: C); (E: A, C); (F: A); (G: B, C) can either be scored by hand or machine scored using an OMR data terminal connected to a microcomputer.

**Linked Question Format**

This format provides a convenient vehicle for assessment of decision-making, which is one of the most difficult areas of student performance to address. Extended-answer questions are typically used for this purpose, such items are difficult to grade objectively and are very time-consuming to grade when large numbers of students are being tested. The basic idea in a linked question is to measures success in matching supporting arguments with initial choices. The format lends itself to individual assessment, group project assessment, or even for laboratory skills assessment. Multiple correct answers are common, permitting awarding of partial credit. The usual caveat of avoiding awarding
negative points for an item applies here as well as with other multiple-response, penalty-scored items. As with other objective-response formats, machine-scoring is effective using OMR data terminals interfaced with microcomputers. (Stand-alone scanners do not have the capability to handle complicated scoring algorithms.)

This format is illustrated below using an example that deals with the purification of water. Several features of linked questions are illustrated by this example. The initial question requires the student to choose a position, and this choice is not graded. In the next step, the student considers data—some of which supports his or her position, and some of which does not. The student’s task is to identify the data that supports the chosen position. The design of these questions should include several supporting statements for each option, as well as some statements that are factually correct but unrelated to the decision. If machine scoring is not available, the item can be easily hand-scored.

The chlorination of water offers advantages and disadvantages. Do you believe that the advantages outweigh the disadvantages?

(A) Yes  (B) No

Choose all factors that provide support for your position.

(C) Chlorination kills most water-borne bacteria.
(D) Heavy metal ions are removed by chlorination.
(E) Trihalomethanes (THMs) may be produced if chlorine is added to water containing organic matter.
(F) Of the various methods for sanitizing water, chlorination is least expensive.
(G) There is a reduction in tooth decay as a result of chlorination.
(H) Ozone and ultraviolet light can also be used to sanitize water, and are safer than chlorination.

If a student chooses (A), then responses (C) and (F) are correct. If (B) is chosen, then (E) and (H) are correct. (D) is incorrect, and (G) is unrelated. A common scoring formula would award 1/2 credit for each correct response and subtract 1/3 credit for each incorrect one. If there were different numbers of correct responses for the two options, each would need its own scoring formula.

Testing for Conceptual Understanding

Many chemical educators are increasingly concerned that students often learn to solve chemistry problems algorithmically, without mastering the underlying concepts. Assessment of conceptual understanding is a strategy for building the content of an item rather than being an item type itself. It is believed that many students, particularly women and some minorities, may have greater understanding of chemical concepts that they are able to demonstrate on examinations that only stress problem solving. Conversely, it may also be true that some students who succeed with problems find conceptually based questions more difficult.

A conceptual-understanding item presents students with a situation in which they must apply their understanding of concepts and interrelationships among concepts. Such an item attempts to probe student understanding (or lack thereof) that may be hidden by reliance on mathematical algorithms. Typical conceptual items might require a student to interpret diagrams, analyze data, explain facts and phenomena, predict outcomes, compare quantities, design experiments, or demonstrate understanding of the three-dimensional, particulate nature of matter. Here is a question that centers on this last point.

The diagram at right represents water molecules before the process of electrolysis takes place. Which diagram best represents the molecular view of the result?

Conceptual understanding items are as amenable to machine-scoring as any other objective-response item, and work particularly well with the grid format, especially when the cells contain graphics. Exploration of the possibilities (and limitations) of conceptual-understanding items is an active area of investigation among chemical educators, and we expect to see much more attention to such items on ACS exams and elsewhere as the techniques are developed and refined.

Lab Practical Items

Chemistry is an experimental science and many teachers feel strongly that the importance of learning in the laboratory should be conveyed to students—by using laboratory-based assessment tools. The laboratory station model works for testing under some circumstances, but it is difficult to implement with large classes, and is often limited by available space and resources. For these reasons, the Examinations Institute has been developing assessment tasks that use small-
scale chemistry as a way to safely and conveniently assess laboratory learning. For more information on the philosophy and pedagogy of these laboratory assessment items, as well as many tested items ready for use with students, please consult the separate publication Small-Scale Assessment Tasks from the Examinations Institute (expected publication date: Late Fall, 1995).

Section 3. How to Gather and Interpret Test Statistics

Why Use Statistics?

Exercising care in constructing assessment items is not enough to assure that the item will prove to be satisfactory on an examination. The item should be tested with real students and the results analyzed. Item statistics are powerful tools for improving the quality of chemistry examinations, and they complement the overall test statistics that many teachers routinely obtain. In this section, we consider the standard measures for test statistics and individual item statistics for normative-referenced tests. These are measures of greatest use to most teachers.

Test Statistics

To begin with, well-constructed tests should be valid; that is, they should measure those student attributes the examiner wishes to assess. There are many aspects to validity, and most are difficult to quantify. Most teachers accept the premise that experts in the field are able to establish face validity satisfactorily. This does mean that we should be able to tell a professional colleague what we intended to measure, have them inspect the examination carefully, and tell us whether or not they agree that our examination measured those attributes.

After administration of an examination, a useful beginning for data analysis is to construct a plot of raw score against the number of students earning each score. This example is from a recent ACS examination in organic chemistry.

![Histogram of Raw Scores](image)

If the histogram is skewed toward either end, student scores in the compressed range do not effectively differentiate among knowledge levels. Occasionally, two peaks are seen in a histogram, suggesting that there are two distinct populations in the class (such as both remedial and advanced students). In addition to determining whether the pattern of distribution of scores is reasonable, the examiner needs to know whether the test was of reasonable difficulty, whether it differentiates students who have command of the material from those who do not, and whether the statistics can be trusted. The calculation of mean, median, and standard deviation are familiar measures, as are percentile rankings. These measures speak to the difficulty of the test overall, but do not address the other issues.

The mean and the standard deviation are both measures of the central tendency in a distribution. The mean score is calculated by adding all of the scores in a set and dividing this sum by the total number of scores. It is the familiar “average” score. Looking at the mean value will give you a quick idea of how students have performed. A very high mean relative to the number of possible points would be an indication of an easy test, but a low mean does not always indicate a difficult test. It may reflect that the test was not valid for the students to whom it was given.

The standard deviation is a measure of the variability of scores from the mean. A relatively small value for standard deviation indicates that scores are tightly clustered about the mean. A large value shows scores to be spread out over a larger range. Standard deviation is calculated by first computing the deviation for each person’s score from the mean score for the sample. These deviations are then squared, and divided by the number of test scores. This process yields the quantity called the variance, and its square root is called the standard deviation, given the symbol $s$.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

$X$ is the individual test score

$\bar{X}$ is the mean test score

$\Sigma$ is the summation

$n$ is the number of scores

The median is the middle score and therefore half the students score above the median and half below it. If the median differs significantly from the mean, the distribution of student scores is skewed—compressed at either the high end (if the median is above the mean) or the low end (if below the mean). The distribution of scores can also be presented as a histogram, such as the one shown on page 16 for a recent ACS examination.

Percentile ranks are calculated by summing the number of students in the sample who achieve up to any particular score and then dividing by the number of students. For example, if a raw score of 55 corresponds to the 77th percentile, this means that 77% of the students who took this test had a score of less than 55, but 23% of the students had a higher score.

Tests should also be reliable, a term that refers to the internal consistency of test scores. One measure of reliability is called the KR21 reliability estimate, which considers whether students would earn the same score if an equivalent examination were taken. It is usually not practical to give students exactly equivalent examinations twice, and internal consistency methods are often used because they require only a single administration of a test. The KR21 reliability
estimate is one of the most widely used formulas for measuring internal consistency. The formula allows calculation of test reliability from three pieces of information—the number of items on a test, the mean value, and the standard deviation.

\[
KR21 = 1 - \frac{\bar{X}(n - \bar{X})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}
\]

\(\bar{X}\) is the mean test score
\(s\) is the standard deviation
\(n\) is the number of test items

If there were a complete lack of correlation between scores on "equivalent" exams, the reliability coefficient would have a value of zero. The limiting upper value is 1.00. The reliability for standardized ACS tests is frequently about 0.90 or greater, although classroom tests might be expected to range in reliability between 0.60 and 0.80 using this formula. If questions vary significantly in difficulty, this particular formula tends to underestimate reliability, so the value of 0.90 for a particular test is, if anything, less than the true reliability.

The standard error of measurement (SEM) expresses the uncertainty in the measurement of an individual test score. It therefore gives a way to interpret the uncertainty in any student's individual test score. This is the formula used.

\[
SEM = s\sqrt{1 - KR21}
\]

\(s\) is the standard deviation
KR21 is the reliability estimate

The value of this parameter tells you that if a particular student receives a score of 52 on a test that has a standard error of measurement of 4.20, then you could be reasonably sure that the range of scores 48–56 would actually include that student's "true" score. The size of the scoring band reminds us not to overinterpret test results, for a student scoring 52 is not proven superior to one scoring 50, due to the overlap of their scoring bands in this case.

Another factor to remember about reliability and its measurement is that it is an important indicator of the value of the test as applied to a particular group of students, rather than an attribute of only the test itself. If an un instructed group of examinees were given this same test, the reliability would be much lower. This is why, for example, a separate standardized test was developed for students in the ACS Chemistry in the Community curriculum, rather than suggesting that they be given the regular high school chemistry test. Course objectives must match test objectives to achieve high reliability.

Item Statistics

There are three types of item statistics that are relatively easy to obtain that are useful to teachers. These are the difficulty index, the discrimination index, and the distribution of incorrect responses. The difficulty index, as used by the Examinations Institute, is simply the percentage of students who responded correctly to an item. The symbol used is \(p\)—the proportion of correct answers on the item. If 350 of 500 students successfully answered any given question, the difficulty would be reported as 70.0. If only 150 of 500 students were successful, the difficulty would be 30.0.

\[
p = \frac{\sum_{i=1}^{n} a_i}{N} \times 100
\]

\(p\) is the difficulty index
\(N\) is the number of students
\(a_i\) is the score for each student

The lower mathematical limit of \(p\) is 0.0, which occurs if not even one student is successful on the item. The upper limit is 100 if all students correctly answer an item. The larger the value of the difficulty index, the easier the question. Items with either very high or very low proportions of success are usually eliminated from a standardized test by the Examinations Institute because such items obscure differences in levels of achievement among the students tested. Items with \(p\) factors greater than 70 or less than 30 are usually rejected, although you might decide to use such items in your classroom testing to achieve other objectives.

For the purposes of designing a standardized achievement or diagnostic test, items that are either too easy or too hard also provide limited information to the examiner because they lack discrimination.

The discrimination index measures the performance on the item for students who did well on the test overall relative to those who did poorly. This index provides a mathematical expression for the intuitive understanding that a well-designed question should result in greater success for the top students than for those students whose overall achievement is lower. For example, if all of the top students answered the question correctly, and none of the low group students did so, the discrimination index would be at its maximum. Conversely, if all of the low-achieving students answered this question correctly, but all of the high-achieving students missed it, there must be something seriously wrong with the question. There may be a widespread misconception on the part of the students. Or, there may have been an error in marking the key!

As commonly used, the discrimination index, here represented by the symbol \(r_s\), is calculated by subtracting the number of right answers on that item given by students in the bottom 27% of achievers from the number of right answers given by students in the top 27% of achievers, and dividing by the number of students in one of these groups. Sometimes groups of 25% or 33% are used, but the upper and lower 27% is considered to be the best compromise between the conflicting needs to make the comparison groups both as different as possible and still as large as possible.

\[
r_s = \frac{n_{\text{high group}} - n_{\text{low group}}}{n_{\text{group}}}
\]

\(r_s\) is the discrimination index
\(n_{\text{high group}}\) is the number of students in the top 27% of achievers
\(n_{\text{low group}}\) is the number of students in the bottom 27% of achievers
\(n_{\text{group}}\) is the total number of students on the test

The discrimination index reveals the effectiveness of an item in differentiating among the achievement of the students who are most successful overall and those who are least successful. The higher the value of \(r_s\), the greater the usefulness of the question in differentiating "good" and "poor" students. If the difficulty of the item is either 0 or 1.0, the discrimination index is zero and no
discrimination can take place. If the difficulty of an item is around 0.5, the discrimination index can range from -1.00 and +1.00. Values of $r$ in the range of 0.30–0.50 are typical for items in most standardized tests of the Examinations Institute, with items having values below 0.20 usually having been discarded after the trial test.

The distribution of incorrect responses is another useful indication of item effectiveness. This is most simply reported as a percentage, showing the relative frequency with which each alternative response has been chosen. A close inspection of incorrect response patterns often uncovers the reason for low discriminating power. Ideally, each of the distractors should attract some students, particularly students in the low-achieving group. Much time must be spent in carefully crafting incorrect responses so they are actually plausible, result from logic associated with common misconceptions, and provide viable choices for students. They must never, of course, confuse or trap any student, just correctly reveal the areas of incomplete understanding or processing of data.

Here is an example of a question used on an ACS trial test. Item statistics helped the test committee reach the decision not to use the question on the standardized test, and the patterns are here analyzed to show why the question is not suitable.

| In the titration of iron(II) sulfate, FeSO$_4$ with acidified potassium permanganate, KMnO$_4$, the pink color that signals the end point is caused by |
|---|---|---|---|
| (A) phenolphthalein | (B) MnO$_4^-$ | (C) MnO$_4^-$ | (D) Fe$^{3+}$ |

The difficulty index was 29.0, the discrimination index was 0.01, 54% chose (A), 29% chose (B), 9% chose (C), and 8% chose (D). There were 333 students who attempted this question and the correct answer is (B). Students appear to have been misled by their remembrance from acid-base chemistry that phenolphthalein sometimes makes a solution appear pink.

**In Summary**

Test statistics can be very useful in analyzing your own tests. Gathering data and calculating some common statistical measures is a task made easier with the use of computer hardware and supporting software, and can help you to better assess the areas of strengths and weakness in your students' knowledge. You can also determine whether your objective tests are really evaluating what you intended to measure. Effective assessment is always a necessary early step in improving the quality of instruction.