George M. Bodner
Purdue University
West Lafayette, IN 47907

Statistical Analysis of Multiple-Choice Exams

Disraeli is often quoted as the source of a statement which appears to possess an inordinate attraction for many, regardless of its validity. He is reputed to have suggested the existence of three types of lies, in order of increasing severity, these are "lies, damned lies, and statistics."

The use of a multiple-choice format for hour exams at many institutions leads to a deluge of more or less significant statistical data which are unfortunately all too often either neglected or completely ignored. We will try to present an introduction to certain of the more common words or phrases which are encountered in the analysis of test results, so that these data may become more meaningful, and perhaps more useful as well. We feel obligated to note in passing that rare indeed is the academic discipline which cannot be accused of sharing Humpty Dumpty's claim that "a word means exactly what we choose it to mean, neither more nor less."

Analysis of the Mid-Point

We might best begin with a limited number of definitions of measures of the mid-point of a normal, Gaussian, or bell-shaped distribution of grades. The *mode,* or *modal point,* is the score or scores obtained by the largest number of students. The *median* is the score obtained by the middle student in the group, the score such that half of the students did better, and half did worse. The *mean,* $\bar{x}$, is the sum of the various test scores, x_i, divided by the number of students taking the exam, n.

$$\bar{x} = \frac{\sum_i (x_i)}{n}$$

(The mean is the quantity which was once called the average before the term average came to connote "normal," and therefore became a pejorative term.) The mean is simultaneously the most tedious of these quantities to calculate and the most representative measure of the mid-point of a test distribution.

Distribution of Scores

The simplest measure of the distribution of scores around the mean is the *range* of scores, or the difference between the highest and lowest scores, plus one. A better measure of the distribution of scores is the variance or standard deviation. The *variance,* σ^2, is the sum of the squares of the deviations of individual test scores (x_i) from the mean ($\bar{x}$), divided by the number of scores (n).

$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$

The *standard deviation,* σ, is simply the square root of the variance. Although the number of students enrolled in introductory chemistry classes at some institutions often appears infinite, or at least transfinite, it is usually better to calculate the variance, s^2, and standard deviation, s, in terms of the number of degrees of freedom available in their determination, $n - 1$.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{(n - 1)}$$

Finally, the standard deviation, s, can be determined more rapidly if the variance, s^2, is calculated using either of the following formulas:

$$s^2 = \frac{\sum_i (x_i)^2 - \left(\sum_i x_i\right)^2 \Big/ n}{(n - 1)} = \frac{\sum_i (x_i)^2 - n\bar{x}^2}{(n - 1)}$$

Under idealized conditions, if the distribution of scores were truly Gaussian, 34.13% of the student's scores would fall between the mean and the mean plus one standard deviation, or between $\bar{x}$ and $\bar{x} - s$. Therefore, 68.26% of the scores fall in the range of $\bar{x} \pm s$. In this idealized distribution, 13.59% of the scores would fall between one and two standard deviations above the mean, or between one and two standard deviations below the mean. Only 2.14% of the scores would fall between two and three standard deviations above (or below) the mean, and a total of 99.72% of the scores would fall within a range of six standard deviations around the mean.

Calculation of Scaled Scores

Since the absolute or raw score on an exam does not indicate a student's performance relative to that student's peers, scaled scores are often calculated which unambiguously indicate the student's location within the distribution of scores. Two of the more popular scaled scores are the so-called z- and T-scores. The *z-score* is equal to the number of standard deviations that a student's raw score falls either above or below the mean. For example, if a student obtains a raw score of 15 on an exam with a mean of 45 and a standard deviation of 15, the raw score is exactly two standard deviations below the mean, and the z-score would be −2.00. A raw score of 90 on the same exam would correspond to a z-score of 3.00.

T-scores correspond to a scale on which the mean has been arbitrarily adjusted to 50, and where the standard deviation has been scaled to exactly 10 points. T-scores may be calculated from the raw score (x_i), the mean ($\bar{x}$), and the standard deviation (s), using the following equation:

$$T = 10\left(\frac{x_i - \bar{x}}{s}\right) + 50$$

Our student who was two standard deviations below the mean would have a T-score of 30, whereas the student who was three standard deviations above the mean would have a T-score of 80.

Advantages of Scaled Scores

There are several advantages to z- or T-score data which make these scales attractive. First, and foremost, the students know where they stand in the course at all times. Students who are told their raw scores, the mean, and the range of scores, cannot always interpret these data correctly. Some are overly confident; others are unduly afraid of failure. Using scaled scores, the students know where they stand in comparison with their peers. If the students are also informed of the typical distribution of grades, they can obtain an even better estimate of their standings in the course.

Scaled scores also allow the instructor to add any number of exam scores in the final analysis of grades without worrying about anomalous weighting of one or more of these exams. Regardless of the mean or standard deviation on a given exam, the z- or T-scores can be combined to produce a total which reflects the student's performance on each exam equally. If one wishes to drop the lowest exam score during the final analysis, it seems better to drop the lowest z- or T-score than the lowest raw score. Alternatively if one wishes to weight one exam more heavily than another, all one need do is multiply the scaled score by an appropriate constant.

A third advantage of scaled scores is the ease, and perhaps the accuracy, with which exam grades can be prorated. If a student misses an exam for a legitimate reason, the instructor can prorate that exam by simply averaging the scaled scores on the student's other exams.

Assignment of Grades Using Scaled Scores

Most introductory courses in chemistry are primarily norm-referenced courses in which grades are distributed on the basis of some normal distribution, rather than on the basis of whether the students have met some arbitrary set of criteria. Although the proportion of A, B, C, etc., grades may differ from semester to semester, and from instructor to instructor, one often has some idea what the final distribution of grades will resemble. *T*- or *z*-scores are ideally suited to the first-order assignment of grades in norm-referenced courses. The data in the table illustrate the approximate percentile ranks of normalized *T*-scores. If one wishes to give approximately 10% A grades, for example, one can begin by selecting all *T*-score averages equal to or above 63. Having used the *T*-score distribution to obtain a rough estimate of the final grade distribution, one can then examine individual students near the borderlines, as we usually do, and adjust the curve to suit one's purposes.

Correlation between T-Scores and Percentile Rank for an Idealized Gaussian Distribution of Test Scores

T-Score	Rank	T-Score	Rank
70	97.7	49	46.0
69	97.1	48	42.1
68	96.4	47	38.2
67	95.5	46	34.5
66	94.5	45	30.9
65	93.3	44	27.4
64	91.9	43	24.2
63	90.3	42	21.2
62	88.5	41	18.4
61	86.4	40	15.9
60	84.1	39	13.6
59	81.6	38	11.5
58	78.8	37	9.7
57	75.8	36	8.1
56	72.6	35	6.7
55	69.2	34	5.5
54	65.5	33	4.5
53	61.8	32	3.6
52	57.9	31	2.9
51	54.0	30	2.3
50	50.0		

Rough Estimates of Test Reliability

One of the major advantages of the multiple-choice format is the ability to calculate a variety of data which pertain to the quality, or perhaps the reliability, of the exam, the extent to which the exam discriminates between "good" and "poor" students.

One of the simplest measures of the quality of an exam involves comparing the range of scores to the standard deviation. In general, as the ratio of the range to the standard deviation increases, the test becomes better at discriminating between students of differing levels of ability. For various reasons, the optimal ratio of the range to the standard deviation depends upon the number of students enrolled in the course.

Number of students in course	Optimal number of standard deviations within the range
25	3.9
50	4.5
100	5.0
500	6.1
1000	6.5

Pragmatically we have found that ratios of 5–5.5 for 700–1000-student classes can be obtained readily. Ratios which are significantly smaller would suggest that the exam might not discriminate between students to the extent desired.

The quality, or reliability, of an exam is also reflected by the *standard error of measurement* which represents an attempt at estimating the error involved in the measurement of a student's grade with a particular exam. In theory, the observed score on an exam should lie within one standard error of measurement of the student's "true" score more than two-thirds of the time. As one might expect, the size of the standard error of measurement tends to reflect the number of points on the exam. Therefore one of the easiest ways to interpret this quantity is to compare the standard error to the range of exam scores. Ideally, the ratio of the range to the standard error should be on the order of 10:1, or greater. There are two points where relatively large values of the standard error of measurement become particularly meaningful: (1) when the mean is relatively low, and therefore the standard error is a significant fraction of the student's score, and (2) when the total of the standard errors for several examinations equals or exceeds the difference between grade divisions for the course. Either situation would suggest that the grade assigned to an individual student is perhaps more arbitrary than we might like to admit.

Item Analysis

Information about the quality of an exam is useless if this knowledge cannot be translated into a means for improving subsequent exams. Fortunately, there are data which can be calculated during the analysis of a multiple-choice exam which can provide hints as to how an exam can be improved.

There are two factors which affect the ability of an exam to discriminate between levels of student ability: (1) the quality of individual test items, and (2) the number of test items. The parameters that are particularly useful in analyzing the quality of an individual test question include: (1) the proportion of the students who choose a particular answer to the question, and (2) the correlation between the probability of a student choosing one of the alternative answers to a question and the student's total score on the exam. These parameters are often grouped together under the title *item analysis.*

Analysis of the proportion of students selecting each of the alternate answers to a question provides information on the difficulty of the question, as well as the extent to which answers which were meant to distract students actually functioned as distractors. These data do not indicate whether a question is good or bad, per se. They do, however, allow one to determine whether questions that one feels are trivial are truly trivial, or whether a question is difficult or truly impossible. It has been suggested that questions which are answered correctly by more than 80%, or less than 25%, of the students are of questionable validity. Data on the frequency of selection of wrong answers are useful as well. These data are most useful in revising questions for future use, since they provide a means for probing the attractiveness of distractors which were included to catch the weaker students.

The correlation between the probability of a student choosing a particular answer to a question and the student's score on the exam can provide useful information on the ability of that question to select between "good" and "poor" students. In theory, the student who answers a given question correctly should have a tendency to perform better on the total examination than a student who answers the same question incorrectly. We therefore expect a positive correlation between the probability of a student getting a question right and the student's score on the exam. When the correlation coefficient

[1] Kuder, G. F., and Richardson, M. W., *Psychometrika,* **2,** 151, (1937). For more detailed information on this, or other, means of analyzing multiple-choice exams see *Essentials of Educational Measurement,* by Robert L. Ebel, Prentice-Hall, Publishers, Inc., **1972.**

for a *correct* answer is negative, something is drastically wrong with the question. Either the wrong answer has been entered into the grading key, or the question is grossly misleading. Conversely, we should expect a negative correlation between the probability of selecting a wrong answer and the total score on the exam. The correlation coefficient for wrong answers should therefore be negative, and the occurrence of a positive correlation is somewhat disconcerting.

Questions for which the correlation coefficient for correct answers are between 0.00 and 0.19 are called inferior, or zero-order, discriminators, and should be removed from future exams. Questions for which the correlation coefficients are between 0.20 and 0.39 are good, or +1, discriminators. We have found that most of the questions written by faculty in our department fall within this range. Questions for which the correlation is between 0.40 and 0.59 are very good, or +2, discriminators, and questions for which this correlation is above 0.60, the +3 discriminators, should be bronzed.

The ideal exam would seem to be composed of questions which lead to the selection of each alternative answer by a finite proportion of the student body, with a correlation between the correct answers and the total score on the order of 0.4 or better, and with negative correlations between the most popular wrong answers and the total score.

Coefficients of Reliability

There are a number of statistical formulas for quantitatively estimating the reliability of an exam, in addition to the rough estimates of test reliability discussed previously. The *Kuder-Richardson formula 20* (*KR-20*), for example, calculates a reliability coefficient based on the number of test items (k), the proportion of the responses to a test item which are correct (p), the proportion of responses which are incorrect (q), and the variance (σ^2 or s^2).[1]

$$r = \frac{k}{(k-1)}\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

This formula cannot be applied when the multiple-choice questions involve partial credit, and it requires a detailed item analysis for calculation. Of the numerous Kuder-Richardson formulas, a second, known as formula 21, has attained some popularity. The *KR-21* reliability coefficient is calculated from the number of test items (k), the mean ($\bar{x}$), and the variance (σ^2 or s^2).

$$r = \frac{k}{(k-1)}\left(1 - \frac{\bar{x}(k-\bar{x})}{k\sigma^2}\right)$$

This formula has the advantage that item analysis data are not included in its calculation; unfortunately, this formula severely underestimates the reliability of an exam unless all questions have approximately the same level of difficulty.

Before answering the obvious, and pragmatic, question of the significance of a *KR-20* coefficient of, for example, 0.697, we must first note that regardless of the ability of an individual test question to select between the "better" and "worse" students, as the number of test questions increases, the level of discrimination increases as well. Therefore, the reliability coefficient for a 20-item test cannot be compared directly with the same coefficient for a 50-item test. Fortunately, using the *Spearman-Brown prophecy formula,* we can predict the reliability of an exam which is made n-times longer (r_{new}) from the reliability of the shorter exam (r_{old}) and the value of n.

$$r_{new} = \frac{n \cdot r_{old}}{(n-1)r_{old} + 1}$$

If the *KR-20* coefficient for a 25-item test is 0.697, doubling the number of test questions should increase the reliability coefficient to 0.821, assuming that all items discriminate neither better nor worse than the first 25.

To study the significance of the magnitude of the *KR-20* coefficient we have examined a set of 51 general chemistry exams used at Purdue University within the last few semesters. This set contained examples from all levels of our program, from the most remedial to the most advanced courses. In each case the Spearman-Brown prophecy formula was used to predict what the coefficient would have been if the test had contained 50 questions. The mean value for the predicted *KR-20* coefficient was 0.779, and the standard deviation was 0.076. It was interesting to note that the reliability coefficient was more susceptible to changes in instructor than to changes in the course to which an instructor was assigned.

It should be noted that the standard error of measurement, discussed previously, can be calculated from the standard deviation for the exam (σ or s) and the reliability coefficient (r).

$$\mathrm{SE_m} = \sigma(\sqrt{1-r})$$

Cum Grano Salis

Under certain circumstances, the statistical analysis discussed here begins to resemble an introductory chemistry student with a Texas Instrument calculator; both provide answers to thirteen significant figures, all of which may be wrong. Under what conditions are we advised to accept these data with a grain of salt?

It appears that item analysis data for individual questions are valid, regardless of the number of questions, so long as the number of students taking the exam is sufficiently large, i.e., on the order of 100 or more. Attempts to apply this technique to studies of differences between sections of 24 students in a multi-section course led to totally unreasonable results.

The nature of the assumptions behind the Spearman-Brown prophecy formula make calculations of reliability coefficients for exams which include only a very limited number of test items worthless, regardless of how many students take the exam. It is our opinion that these data are meaningful for exams which include a minimum of at least 15 test questions.

Conclusions

There are at least three advantages to the use of a multiple-choice format for exams in courses which contain a reasonably large number of students. First, and foremost, we have found that careful consideration of the results of item analysis can lead to significant improvements in the quality of exams written by an instructor. Second, the multiple-choice format provides a consistency in grading that cannot be achieved when exams are graded by hand. Third, the use of the multiple-choice format for at least a portion of each exam frees teaching assistants and faculty for more pleasant, as well as more important, tasks than grading exams.

✦ ✦ ✦