EFFICIENT VALIDATION OF TEACHING AND LEARNING USING MULTIPLE-CHOICE EXAMS

Mark DeSantis and Thomas A. McKean

Department of Biological Sciences and WWAMI Medical Education Program, University of Idaho, Moscow, Idaho 83844

One purpose of this study was to quantify, by means of single-format, multiple-choice questions at the beginning and end of the course, the extent to which first-year medical students learn neuroscience material from an introductory course in their curriculum. Compared with their precourse test performance (mean = 41.8%), collectively, the students nearly doubled their grade by the end of the course (mean = 81.4%). Their scores in subcategories of the material improved in inverse proportion to what they knew initially. A second goal was to evaluate a two-dimensional, computer-generated matrix as a way to assess test question validity and value. The evaluation of individual test questions as assessed from the matrix often, but not always, was similar to the classical pedagogical analysis that uses difficulty and discrimination indexes. Strengths of the matrix are its ability to render data as a gestalt, as well as flexibility and intuitive ease of use.


Key words: course examinations; test question analysis; matrix; medical; neuroscience

Certain aspects of education are relatively invariant over long periods of time; others are quite variable. That dichotomy can exert pressures on the educational process. For example, a constant in medical education is the responsibility instructors have to their students and the public to verify a standard of satisfactory performance by those who have successfully completed a segment of training (1). That obligation for accountability can be buffeted by things that are often in a state of flux, including the instability associated with curricular change and greater demands placed on a student’s and/or instructor’s time (13, 21). Other examples where there may be high variability include the entry-level knowledge students have in a particular discipline and the accretion of new knowledge in a field of study (7, 16, 18). Such demands may make it of practical importance for the instructor to modify courses accordingly. Examples of changes in the way medical school courses are delivered have been reported (6, 14, 19).

When such varied educational situations go on in an overall environment that has an increased pace, it is important for both instructor and student to be able to carry out their tasks efficiently and effectively. For instructors, that responsibility applies not only to delivery of content but also to validation of student performance by effective testing. Question quality is an important determinant in testing (2), and methods have been described for recognizing and handling troublesome test items (11). Fortunately, technology, which itself produces some of the pressures alluded to above, can also prove advantageous in resolving confrontations between invariant and variable aspects of the educational process.
We tested the hypothesis that students actually do improve their exam performance in neuroscience after taking a course about that topic. We did so in a way that provides several advantages, including a more complete record against which future course changes may be benchmarked. Some of these data have been presented as an abstract (4). Second, we describe a new, intuitive method for evaluating multiple-choice questions on the basis of student performance that involves the construction of an ordered array using readily available computer software.

**METHODS**

**Course description.** There are six campuses in five states (Washington, Wyoming, Alaska, Montana, and Idaho, known by the acronym WWAMI) for teaching first year students of the University of Washington School of Medicine (UWSM) in Seattle, Washington (15). At the end of a course, we assess student performance at all of the sites with an identical group of test questions as one way to standardize to some degree the same course being taught at multiple sites by different instructors. That set of common questions constitutes the examination referred to in *Post-course test.*

Medical students at two of the six campuses, University of Idaho (UI) and Washington State University (WSU), which are eight miles apart in Moscow, Idaho, and Pullman, Washington, take classes together and are the study group of this report. The majority who take this neuroscience course at UI and WSU are medical students (usually 36/yr); however, three graduate students were enrolled in the course during the 3-yr period of study in the late 1990s. Results were tabulated for all 111 (34 + 38 + 39) students in those years. They were not made aware that they were part of this study. The mean undergraduate grade point averages for the three sequential classes of WSU/UI medical students were 3.59, 3.68, and 3.71. Their respective average scores on the Medical College Admission Test (total for physical science, biological science, verbal reasoning, and writing sample) were 35.82, 36.21, and 37.98 (J. Carline, personal communication).

The five-credit-hour, one-semester (January through May) course of interest for this study is an interdisciplinary introduction to neuroscience that is part of the 1st-yr curriculum of the UWSM. The course is taught in a conventional lecture (37 sessions, 50 min each) and laboratory (13 sessions, 2-3 h each) style that is augmented by small group sessions dealing with problem solving using neurological case histories (3 sessions, 2 h each) and a website (http://www.sci.uidaho.edu/med532/). Lectures were formal presentations by the faculty; students often did ask questions during and just before/after the session. Laboratory periods involved using a syllabus, microfiche, and gross specimens of the central nervous system. Students worked in self-selected groups of two to five during lab, and they often asked questions of the three faculty members who were present. Optional (no grading) quizzes with the use of projected slide images were irregularly included toward the end of some lab periods. No attendance records for lecture or lab sessions were kept. The total number of course hours, 76, places it at the median range reported for neuroscience courses at allopathic and osteopathic medical schools in the United States (5).

In addition to the pre- and postcourse tests described in the following sections, there were three lecture/lab test sets given during the course that were prepared solely by the faculty teaching the UI and WSU students (Fig. 1). New content tended to build on topics covered earlier in the course, but for the three in-course exam sets, the focus was on material covered since the previous one (or from the beginning of the course in the case of the initial test set). Those six in-course tests were weighted equally and constituted four-fifths of the overall course grade, with the postcourse or “common questions” test results adding the remaining one-fifth (Fig. 1). Requirements for earning a passing grade for the course were an overall course grade of ≥70% as well as ≥70% on the postcourse test.

**Precourse test.** At the first class session, students were given eight questions, written by one of us, to answer anonymously (i.e., no name on the test) over a period of 15 min (Fig. 1). Students were told that those questions covered material that would be taught throughout the course and that the format of questions was the same as would be used in lecture tests later in the course, including the postcourse test. The format is a standard one. An assertion or query is followed by five alternative answers from which the
Sequence and valuation for tests given in the course. The height of each filled bar gives the test’s proportional contribution to a student’s overall course grade. A precourse test was given the 1st class day; it was taken anonymously and therefore did not contribute to a student’s grade. A postcourse test administered the last day of the course constituted 20% of the overall grade. During the course, there were 3 sets of tests, each set having lecture (multiple-choice questions) and lab (fill in the blank) components counting equally toward the remaining 80% of the overall course grade.

This test was also given to four seniors in high school who had no formal training in neuroscience. They were participants in a summer program for students judged by their teachers to be academically competitive candidates for careers in the life sciences.

**Postcourse test.** During the final-exam week, students enrolled in the course took a test composed of 50 multiple-choice questions (Fig. 1). It covered material that had been taught throughout the course. These test items were written, edited, and approved by the WWAMI faculty at the six universities who participate as instructors in this course. They considered this composite to be core material on which every student who took the course should be able to earn a score of ≥70%. These tests were permanently collected and graded, and an item analysis was done on each question.

**Data analysis.** A two-way analysis of variance (ANOVA) was done to test for differences in scores due to pre- vs. postcourse tests and years and due to an interaction between these factors. For all tests, the significance level was α = 0.05. Also, student performance on subcategories of questions was tested for statistical significance (P < 0.05) using the Pearson correlation coefficient after regression analysis was done. Either the average postcourse grade minus the average precourse grade (i.e., amount of change in performance) or the average postcourse grade itself was plotted against the average precourse grade. The same questions, given in the same order, were used on both pre- and postcourse tests for each year. Measurements are expressed as means ± SD, and units are in percentages, with a maximal value being 100%.

A spreadsheet was set up using conventionally available software (e.g., Microsoft) to do the matrix analysis of each student’s overall grade for the course arrayed against difficulty level for 30 of the postcourse test questions. In addition, the difficulty index (p) and the discrimination index (D) for each of those 30 questions were calculated using standard formulas (9, 10, 12). The value p represents the proportion of students who answered the question correctly of the number who attempted the item; it is a value between 0 and 1. A standard interpretation of p is that values >0.75 designate relatively easy questions and values <0.25 indicate more difficult ones (9). The value D was calculated as the number of students with an overall grade in the top half of the test takers who answered the question correctly minus the number of students with scores in the bottom half of the group also answering correctly divided by the number of students in the larger of these two groups. The range for D is from +1.0 to −1.0, and it is evaluated as to both strength and direction of its discriminative ability. Values in the range from +0.2 to +0.4 are indicative of items with positive, moderate discrimination, whereas values in the range from +0.2 to −0.2 are considered to be either positive or negative, respectively, but weak discriminators (9).

**RESULTS**

**Average test scores.** The postcourse average test score was significantly better (F = 509.23, P > F = 0.0001) than that for the precourse test (Fig. 2). On
the precourse test, the average grade over all 3 yr was 41.8 ± 5.5% (35.5, 44.8, 45.2%). Collectively, the average grade for the precourse test was about twice the score that would be expected at a purely chance level for answering correctly (i.e., 20%). The mean score on the postcourse test was 81.4 ± 3.2% for the 3 yr (78.9, 80.2, 84.9%). After having taken the course, the students again nearly doubled their score (Fig. 2). The respective class averages each year on the pre- and postcourse tests were not significantly different ($F = 0.64, P > F = 0.53$).

The average overall course grades in successive years of the study were 83.2 ± 5.8, 84.8 ± 8.2, and 87.3 ± 9.6%, respectively, for an average of 85.1 ± 2.1%. Thus there was a gradually higher grade in successive years for both the overall course (increases of 1.6 and 2.5%) and the postcourse test or common questions (increases of 1.3 and 4.8%; Fig. 2). In general, a student’s overall course grade, although usually a bit higher than his/her score on the postcourse test, correlated directly with it ($r = 0.8849$).

**Findings from categorical groupings of questions.** None of the eight precourse questions was answered either incorrectly or correctly by all students in the study (Fig. 3A, top). Yet for particular questions, it was possible to detect trends in the proportion of students getting a wrong (or, alternatively, right) answer. For example, questions b, c, and h were answered incorrectly by more than one-half of each class, whereas questions a and f were incorrectly answered each year by less than one-half of the class (Fig. 3A, top). Similarly, for the postcourse test, there were no questions that all students answered correctly or incorrectly. There were 14 questions that students in at least one of the three classes uniformly answered correctly, and three of those (questions 23, 26, 42) were answered correctly by all of the students in 2 of the 3 yr (Fig. 3A, bottom). Conversely, questions 12, 16, and 24 were answered incorrectly by at least one-half of the class in most years of this study.

**FIG. 2.**
Class averages on pre- and postcourse tests are shown for the 3 consecutive years of this study. The average percentages for pre- vs. postcourse tests were significantly different from one another for each yearly comparison. Standard deviations were consistently less for postcourse compared with precourse tests. Arrows on the ordinate represent the average score that would be expected at a chance level of performance (bottom) and the overall grade required to pass the course (top). Nos. of students turning in exams are given within each column; not all students turned in a precourse test in the 2nd and 3rd yr of the study.

**FIG. 3.**
Test item analysis by question and by category. A: all 8 pre- and 50 postcourse test questions are aligned on the respective ordinates from top to bottom. Large dots represent the 3-yr average for the percentage of students answering the question incorrectly (abscissa); small dots are for individual years. Arrow and bracket demarcate, respectively, the average ± SD for the percentage of incorrectly answered questions on the 2 tests over the 3 yr of this study. Overall, there was a smaller percentage of incorrect answers on the postcourse test. Arrowheads along the abscissa delineate postcourse questions arbitrarily into 3 groups based on the percentage of students answering incorrectly. From left to right, we considered the groups of questions to be “easy” (<10% got wrong), “moderate” (10–30% got wrong), and “hard” (>30% got wrong). B: pre- and postcourse test questions were subdivided into categories based on either the sequential organization of the course (top quartet) or the basic science/clinical correlation emphasis (bottom duo). In every type of pre- vs. postcourse comparison, the average percentage of wrong answers was less on the postcourse test, and the difference between the means within each pair is statistically significant.
Figure 3B portrays the data when pre- and postcourse test questions are nested in two ways. In Fig. 3B, top, grouping is on the basis of how a question related to the chronological and systems organization of the course. The arrangement in Fig. 3B, bottom, relates to whether a question’s primary emphasis was basic science or clinical correlation material. Regardless of the criterion for subdividing, student performance was always significantly better on the postcourse questions (Fig. 3B). Interestingly, the point gains in a class’s performance varied inversely with initial test results. When each nested group for each year was evaluated as an entity, the respective Pearson correlation coefficients \( r = 0.836 \) (\( n = 12 \), \( P = 0.0072 \)) and \( r = 0.949 \) (\( n = 6 \), \( P = 0.0038 \)) showed a significant inverse correlation between amount of improvement and initial test score. For example, the largest differences between pre- and postcourse sets of average scores in each nesting were on questions about “introductory material” and “basic science”; those were also the topics in each nesting arrangement where students had the greatest percentage of incorrect answers on the precourse test (Fig. 3B). The less students initially knew about an aspect of neuroscience covered in the course, the greater was their gain by the end of the course. In contrast to that consistency, there was not a significant correlation between pre- and postcourse test performance for the various subcategories in either nesting paradigm (\( r = 0.171 \), \( n = 12 \), \( P = 0.596 \); \( r = 0.483 \), \( n = 6 \), \( P = 0.332 \)). Thus the absolute value for a class’s initial score in a subdivision of the course did not predict well the postcourse test result in that same aspect of the course.

**Analysis of postcourse test questions displayed on a matrix.** The occurrence of incorrect answers was charted on a two-dimensional matrix constructed from each student’s overall grade for the course and test questions of varied difficulty selected from the postcourse test (Fig. 4). Student performance is plotted against question difficulty in a high-to-low order along the respective axes. That array was then examined for patterns among the variables. The questions whose numbers are listed along the top are a subset from the postcourse test. They include all 10 questions missed by >30% of the students, another 10 questions missed by between 10 and 30% of students, and a like number missed by ≤10% of students. To have an equal number of questions in each category and thereby make comparisons easier on visual examination, those in each of the latter two categories were selected by choosing every other question in the respective percentile ranges shown in Fig. 3A, bottom. If one scrambles the values on both the \( x \) - and \( y \)-axes, a control matrix against which to check the ordered one is generated. The scrambling may be done, for example, by arranging the \( x \)-axis on the basis of the chronological order of the question number (i.e., question 1, 3, 4, 5, 6, \ldots 48 along the top of Fig. 4), rather than in a hierarchy of students answering incorrectly. The \( y \)-axis was disordered by ranking students according to an alphabetical listing of their initials (not shown for reasons of privacy) rather than by aligning individuals according to their overall course grade.

The ordered matrix in Fig. 4 shows that the greatest density of “dots” (incorrect answers in this example) is toward the left and top, and fewer dots are in the right and bottom parts (easier questions and higher overall course grades). Questions in each of the three arbitrarily designated difficulty levels contributed to distinguishing student performance, although at different boundaries. For example, no student whose
overall grade was ≥90% missed any of the easiest questions, whereas about one-third (22 of 72) of the other students with an overall passing grade in the course did answer those incorrectly. Nearly two-thirds (3 of 5) of students who had a failing overall grade missed questions in this category, and the four questions that were the very easiest (i.e., questions 19, 23, 26, 38) were missed only by students very close to or below the boundary level for passing the course. Thus a cluster of even the easiest questions helped to identify students at the fail/pass boundary.

If one examines the two sets of more difficult questions (i.e., those missed by >10% of all students), then other patterns that demarcate overall grade boundaries are evident. For example, there are five questions (questions 12, 24, 32, 34, 37) that were invariably missed by each of the five students who earned an overall grade of <70% (Fig. 4). One of those questions (question 24) is not very helpful in distinguishing among the students, because they usually answered it incorrectly regardless of their grade; it represents a defective question (see also below). Another (question 12) helps distinguish among the better-performing students, since no consecutive group of five students that had a grade >92% got it wrong. For the remaining three questions missed by all of the students who failed the course (i.e., questions 32, 34, 37), in no instance does one see a similar contiguous sequence of five students missing them among the students who earned an overall passing grade. Those three questions, then, also help to distinguish students at the fail/pass boundary. On the control or scrambled matrix, there were no questions missed by all five students at either end of the γ-axis (not shown).

By use of a similar but less stringent approach, there are another three questions (questions 29, 43, 47) that were missed by four of five students who had a failing overall course grade (Fig. 4). One of those questions (question 43) was in the hardest category and the other two (questions 29, 47) were in the moderate category for difficulty. For those three questions, there were similar contiguous strings of four of five students answering incorrectly among the passing students. All of those clusters were below the overall mean grade for the course (Fig. 4). This trio of questions did not distinguish at the fail/pass or the pass/honors boundaries, but it did identify students performing in the upper half from those in the lower half of the classes. On the scrambled array used as a control, there were four questions that were missed by four of five students in sequence at either the very top or bottom of the array. However, in each of those four instances, similar contiguous sequences were scattered throughout the scrambled matrix from top to bottom (and from left to right) rather than being grouped within a part of the matrix (not shown).

Yet another way the matrix may be used to identify questions that demarcate boundaries of student performance was to note which questions are missed by at least twice as many students with an overall course grade below the mean as by those above it. In this case, the demarcation is between the upper and lower halves of the range in grades. By that criterion, nearly all of the questions that are bounded by questions 43 and 33, listed along the top of Fig. 4, were in this category (questions 5 and 6 were exceptions). That group of questions constituted nearly all of the ones missed by 5–40% of the students (Fig. 4). On the control matrix, no question met this criterion (not shown).

**Comparison of matrix with p and D indexes for item analysis.** Because the matrix system is being proposed as a new method of test analysis, it is important to see how it relates to the p and D indexes currently in use. That relationship for the 30-question subset from the postcourse test is charted in Fig. 5.

It is clear that the distribution is skewed along the abscissa toward the “easy” side of the p index scale. That means that there is a discrepancy between the matrix assessment of difficulty—in which there were equal numbers of questions in the categories considered hard, moderate, or easy—and the more conventional difficulty index (p). For example, for only one question (question 24) was there agreement between the two methods that the question was hard. However, the order of the questions in Fig. 5 from left to right (harder to easier) is in much the same alignment as that seen along the abscissa of the matrix analysis (Fig. 4).

Along the D index axis (ordinate of Fig. 5), the questions that distributed near the top are moderately to strongly discriminating on the basis of the index anal-
ysis as well as the matrix analysis (circled question nos.). Those points, as well as many of the ones tracking along the slope at the right side of Fig. 5, are questions that were judged on the matrix analysis to distinguish at the lower-from-upper boundary of students’ performances (triangulated question nos.). There was also agreement between the methods that questions 5, 24, and 41 were not discriminative ones.

Apparent inconsistencies between the matrix and the index analyses were as follows. Most of the easiest questions on the $p$ index also had a low $D$ index value, yet many of those same questions, when looked at from a matrix perspective, did help distinguish at fail/pass, lower-half/upper-half, or pass/honors levels. Of the 8 circled questions, 6 were, using classical numerical analysis, of average difficulty, and 7 of the 8 were either moderately or strongly discriminating. The 16 nos. enclosed with triangles were questions that by matrix analysis identified the lower half from the upper half of student performance. According to their index values, slightly less than one-half were of average difficulty, and slightly more than one-half were moderate or strong, positive discriminators.

**FIG. 5.** Comparison of matrix interpretations with difficulty ($p$) and discrimination ($D$) indexes for test items. Each of the 30 postcourse test questions portrayed on the matrix (Fig. 4) is represented by the test item no. plotted according to its $p$ and $D$ index along the $x$- and $y$-coordinates, respectively. Vertical lines on the graph show boundaries at conventionally defined levels of difficulty (hard, average, easy); horizontal lines demarcate standard levels of discrimination (none, moderate, strong). The 8 encircled nos. shown are those that on the matrix were a “contiguous cluster” (Fig. 4). One of the 30 questions (question 24) was hard by both matrix and index methods of analysis and was not a discriminating question by either method of analysis. All other question nos. enclosed within a circle are ones that by matrix analysis were judged to help distinguish at fail/pass, lower-half/upper-half, or pass/honors levels. Of the 8 circled questions, 6 were, using classical numerical analysis, of average difficulty, and 7 of the 8 were either moderately or strongly discriminating. The 16 nos. enclosed with triangles were questions that by matrix analysis identified the lower half from the upper half of student performance. According to their index values, slightly less than one-half were of average difficulty, and slightly more than one-half were moderate or strong, positive discriminators.
and postcourse test scores whereas our results were ~40 points. On the basis of the style of multiple-choice question used by Richardson (Ref. 16, see his Table 1), where three or four choices were presented, the chance level of performance would be ~26.0 ± 2.9%. The precourse test average for his combined groups was only slightly above that value (28.4 ± 3.9%). Thus, unlike the entry scores of the students in the present study, Richardson’s students were performing just above chance level at entry to the upper-division course even though two of the three groups had taken at least one introductory physiology course. A possible reason for why this was so is suggested by the study of Swanson et al. (20), who reported that students showed a “modest decline in retention” of basic science material by three years later (see also Ref. 17).

Our data on pre- vs. postcourse testing showed that students made the greatest relative advances in those topic areas about which they initially knew the least. This may represent a “ceiling effect.” That is, students who score low initially have greater room for improvement in their grade than those who score high initially, when all of the students’ grades end up at a higher, but similar, level by the end of the course. Our findings on score improvement are consistent with those reported by Richardson (16). The gain achieved between pre- and postcourse values by his three groups of students was inversely related to the group’s starting performance level. This notion of greater gain resulting from a lower starting position also tends to account, at least in part, for why medical students entering with varied amounts of science-based undergraduate backgrounds do equally well in their preclinical courses (7). Given the ample motivation and intelligence of medical students, those who may be initially less versed in a particular topic will usually have shown the greater gain at the end of a course. An assumption for this explanation is that the teaching effectiveness on the part of the faculty is roughly constant throughout various components of a course or curriculum. We believe that that was the case in the present study, since most of the faculty for the course have worked together for more than a decade and gave lectures in more than one subdivision of the course as well as participated in all laboratory and case study sessions.

DISCUSSION

Pre- and postcourse test performance comparisons. It was heartening to learn that students improved their test scores in neuroscience after having taken a course in the subject. This study provides a measure of how much improvement occurred. The students’ average precourse test performance was about twice what the expected average score achieved by chance would be (i.e., 20%, given the question format we used). They had another almost twofold improvement in scores at the end of the course. Additionally, the study provides information about a possible underlying principle by which students improve their test scores, at least in the short term, by taking a course.

On the basis of the average performance by high school seniors on the same precourse test (21.9 ± 6.3%, n = 4), which was not different from chance level, the medical students probably gained their entry-level knowledge of neuroscience in the years after high school. By way of comparison, Richardson (16) found that college students improved their score by 50–60 points on a postcourse test in physiology compared with precourse results. The larger improvement (60 points) occurred when the same questions on both tests were compared; different topic questions on the pre- vs. postcourse test resulted in the smaller average gain (50 points). The latter regimen is closer to the one used in the present study, where the average point gain was ~40 points. There was the same number of questions on the precourse test in both studies. Interestingly, there was no difference in improvement among Richardson’s study groups when he compared those having had a previous course in physiology with those that had not (16). All three of his groups started with precourse test scores that were not significantly different among themselves (29.4, 31.7, and 24.1%), but those entry-level scores were all notably lower than the ones for our students. The initial score differences between the two studies can account for why Richardson’s groups showed a 50–60 point improvement between pre-
Another feature of our results was that the standard deviation around the mean score was invariably less on the postcourse test than on the precourse one. There are at least three explanations for why that occurred. The decreased variation on the postcourse test may reflect a standardization of thinking that was brought about by the teaching/learning process. Second, as everyone’s grade in the course gets close to the maximum achievable score, the amount of variation should decrease. Finally, there were about one-sixth as many questions on the precourse test as there were on the postcourse test, so an answer on the former test was worth a larger value (12 points) than on the postcourse test (2 points). Given that the number of students taking both tests was relatively constant, the larger increment between successive grade levels is likely to result in a larger standard deviation about the mean grade.

Test question analysis using a matrix. There are various strategies for analyzing test questions. One way, for example, is to have those who write questions for national exams evaluate test questions written by others (8). A more traditional approach is to make a judgement on the basis of the performance by students who actually have answered the questions. That is the approach that we used. However, we describe a new way for assessing test questions, which employs the advantages of readily available computer software to construct a two-dimensional matrix, which we compared with the more classical pedagogical analysis of discrimination and difficulty indexes. When using the traditional indexes, one assesses the worthiness of a test question depending on the numerical value calculated for each question relative to standard, but arbitrarily set, limits within the total range of those numerical values. Furthermore, the index values can then be compared between studies. For example, the type of multiple-choice question that we used can be classified as a “novel-vignette” type (3). The average difficulty and discrimination index values of our 30 postcourse test questions, \(0.76 \pm 0.22\) and \(0.19 \pm 0.13\), respectively, were comparable to those reported by Case et al. (3) for first-time examinees answering the same style of question on a national exam for medical students \(0.73 \pm 0.14\) and \(0.24 \pm 0.11\), respectively).

Notwithstanding the merits of traditional approaches, we believe that the matrix has several advantages. A major one is the immediate visual gestalt one gets of student performance relative to a set of test questions. The test question set is arranged along one axis, which codes for relative difficulty whereas the other axis codes for student grade. No mathematical calculations are necessary beyond determining each student’s grade and how many of the students answered each question incorrectly (or correctly). A second advantage of the matrix is its flexibility. On the basis of the pattern of incorrect (or correct) answers at various performance levels (e.g., grade boundaries), one can readily detect patterns of performance on various test questions. The flexibility in the system also allows for an internal control simply by scrambling the order on both axes of the matrix. This gives a way to verify whether the student performance at boundary zones is legitimately linked to individual test items. Without a computer, randomizing the matrix would be prohibitively labor intensive.

Perhaps the most striking observation evident from the matrix analysis was that questions judged to be among the easiest in the series can have a very real role in identifying select groups of students. One may be inclined to think of the easiest questions as being of little value, since “everyone” is perceived to answer them correctly. But in this study, 40% of the easy-category test questions did allow for confirmation of those students with the lowest grades. An additional 30% of the easy-category test questions helped to identify students performing in the upper vs. the lower half of the class. Virtually all of those questions would be judged by the discrimination index to be without merit. It is true, of course, that none of those questions identifies the students at the highest grade level. But other test questions did that, and even those questions would be judged by the difficulty index to be at an “average” difficulty level.

APPENDIX

Example of a precourse test question (item b in Fig. 3A) classified in the “introductory material” and “basic science” categories:

Which of the following is an autonomic ganglion that contains cell bodies of postganglionic parasympathetic neurons?

- trigeminal ganglion
- superior cervical ganglion
nodose ganglion (inferior ganglion of the vagus nerve)
celiac ganglion
none of the above is correct

Example of a precourse test question (item g in Fig. 3A) classified in the ‘higher/integrative function’ and ‘clinical correlation’ categories:

A stroke that resulted in what is known as an “expressive aphasia” (also called Broca’s or motor aphasia) in a right-handed person most likely affected which of the following arteries or its branches?
right middle meningeal
left anterior cerebral
right posterior communicating
left middle cerebral
right superior cerebellar

Each student who has taken this neuroscience course over the past few decades, particularly those comprising this study, has our appreciation. It has been fun to guide them through part of their education and to learn from and about them. We thank E. Carter for calculating the difficulty and discrimination indexes of the questions, Drs. S. White and M. Laskowski, our colleagues in teaching and discrimination indexes of the question.


