Assessment of Student Performance in Statistics

KEYWORDS:
Teaching;
Assessment.

Carl J. Huberty
University of Georgia, USA.
e-mail: chuberty@coe.uga.edu

Summary
This article describes assessment and scoring methods that have been used successfully in graduate-level statistics teaching.

◆ INTRODUCTION ◆

Many manuscripts have been written about the art, the skill, the science, the craft, etc., of teaching in general (e.g. Brookfield, 1990; Eble, 1979, 1983; Ericksen, 1985; Gullette, 1984; McKeachie, 1994) and of teaching statistical methods in particular (e.g. Cobb, 1993; Garfield, 1995; Hawkins *et al.* 1992; Moore, 1995; Simpson, 1995). However, assessment of student performance is seldom considered an important dimension of instruction in and of itself. Assessment/evaluation is either not considered at all, or it is considered as an element under 'course management'. In the evaluation of college faculty performance in instruction, seldom too is student assessment given the respect that some (including the present author) think it deserves. For example, Braskamp and Ory (1994) and Centra (1993) do not consider student assessment to much of an extent. There is a fairly recent book by Astin (1993), however, that is devoted to assessment of college student learning.

It is maintained by the present author and others that student assessment can very well drive student learning. The 'drive' pertains to (positive) pressure on student study, encouragement for studying on the part of the student and a 'learn by doing' attitude. It is felt that student grades (that are earned, not awarded) should reflect some naturally existing variability across student performance. Before presenting a particular approach to graduate student assessment in statistics, it should be noted that a number of writings pertaining to statistics assessment have appeared during the 1990s. Most of these have been fairly philosophical in nature – this is *not* to be construed as a criticism. Noteworthy of mention are Garfield (1993), Hubbard (1997), Jolliffe (1991) and, especially, Garfield (1994a,b). Some practical aspects of assessing student performance in statistics are described in a collection of writings edited by Gal and Garfield (1997).

The purpose of this article is to illustrate an approach to assessment of student learning in an introductory graduate-level statistical methods course. A three-type-assessment approach has, over a number of years, resulted in a fair bit of variability in reflected student learning. This approach has been utilized by the author for over 25 years. The approach varies a little across course levels, but the general idea remains fairly constant.

◆ AN ASSESSMENT APPROACH ◆

The learning environment context considered is the teaching of applied statistical methods to graduate-level students in the behavioural sciences. A three-course sequence is considered. This is a sequence in statistical methods from (1) introductory methods (descriptives and correlation, statistical inference and inferential comparison of two means and two proportions) through (2) comparison of multiple means and proportions (including factorial analysis of variance and repeated measures analyses) to (3) multiple regression and correlation analyses.

Class activities

Some of the class activities utilized include the usual lecturing and question-and-discussion sessions. Students are, on occasion, requested to formulate 2–3 questions about concepts covered in class discussions or in the textbook. These questions, prioritized by each student, are either discussed with the entire class or discussed in 'learning groups'. Each learning group is comprised of 3–4 students – I determine the initial group structure (by 'mixing' student areas of study) and then 'rotate' group membership for subsequent sessions. Class time is also spent reviewing 'practice items' as well as administering quizzes and the midterm test.

Student assessments

Three in-class assessments are utilized: four quizzes, a midterm test and a final examination. All three of these types of assessments are of the short-answer variety. Some items (mostly in the quizzes) are multiple-choice items. (I would maintain that most of the multiple-choice items call for more than recall of facts and routine computations.) Other items in the quizzes, the test and the examination are other types of short-answer items – these items call for brief explanations, identifications (sometimes from computer printouts) and definitions.

All quizzes are 10-point assessments and typically take the students about 20–25 minutes to complete. Class time is spent discussing each quiz item immediately after all students turn in their responses. (Discussion at this time has always been the choice of the students, in preference to discussing the items during the next class session or not discussing them at all.) Quiz items vary somewhat from section to section of a given course; changes are made because of content covered and item revision.

The midterm test length varies from course to course with the total number of points ranging from about 40 to about 70. In the most advanced of the three courses, one part of the midterm test is of the take-home variety. Items on the test for a given course are not generally the same from section to section; some items are revised or replaced. In each of the three courses the total number of points on the test remains constant from section to section. About 110 minutes are allowed for test completion. Test results are reviewed during the next class meeting.

The final examination is administered after the last scheduled class session of the term. As with the test, the total number of points for the examination varies somewhat across the three courses (from about 45 to about 75), but remains constant across sections of a given course. There is virtually no time limit for students to take the examination.

Except for the take-home part of the midterm test for the third course, all assessments are of the closed-book and closed-notes variety. There is minimal demand for memorization of formulae except for very 'basic' ones. Composition of the four quizzes, the test and the examination for a given course varies somewhat from section to section. I would argue, however, that difficulty level, time-to-take and content coverage are fairly stable within each course. Copies of the quizzes, the test and the examination are not returned to the students for them to keep. All assessments are, however, made available to students for review and study purposes. Over the years, there has been very little concern expressed by students regarding amount of testing, scoring of items or fairness of assessment.

Two quizzes, equally spaced, are administered before the midterm test and two quizzes are administered between the test and the final examination. Quiz 1 covers the first quarter of the term, Quiz 2 covers the second quarter and the test covers the first half; similarly for Quiz 3 and Quiz 4, with the examination covering the second half.

Measurement characteristics of the quizzes, the midterm tests and the examinations are judged to be acceptable. Specifically, content validity of the three types of scores is judged to be very respectable. The value of the Kuder–Richardson 20 index (of internal consistency) for the tests has ranged from about 0.65 to about 0.85 over the past few classes. It is also assumed that a common scale of measurement is used across classes for the quizzes, for the tests and for the examinations.

Other student activities

In addition to the four quizzes, midterm test and final examination, there are two other activities that are related to the assessment process. One is article critiques. Students are expected to locate published journal articles in which data analysis techniques covered in the course are implemented. It is suggested that students locate articles in their area of study. If they have difficulty locating relevant articles, I have a number of references from previous classes from which current students may select an article. The critiques are fairly structured. Across the three courses, students locate two or three articles; different statistical techniques are used in the sets of critiques.

The second student activity is a research project. There are two choices regarding the type of project: a substantive study using data analysis methods discussed in class meetings, or a methodological study related to, but possibly different from, methods studied during the term. (Very, very few students have opted for the second type.) Structure for the study, including reporting methods for the first type, is provided. Research project reports are turned in near the very end of the term.

It may be noted that, for these two activities, student collaboration is permitted. Two or at most three students may work together on the article critiques and on the research project.

◆ SCORING AND GRADING ◆

Scoring

Student performances that yield numerical assessments are the only ones considered regarding student evaluation. Aspects such as class attendance, class participation and homework are *not* considered in the evaluation process. (On occasion, especially in the first course, homework is collected and reviewed, but not 'scored'.)

Of the four quizzes, each student's lowest score is dropped. For each student the three highest quiz scores are summed to yield a Q score; the maximum attainable Q score is 30. The Q scores for a current class are included in a distribution of scores with the three most recent classes for the course. The arithmetic mean and standard deviation for the complete four-class distribution are used to transform a Q score for the current class to a unit standard score; this is labelled z_Q.

The midterm test score, T, for the current students is included with the T scores for the three previous classes for the course. The mean and standard deviation for the four classes are used to yield a z_T score for each student in the current class. Similarly, a z_E score is calculated for each student in the current class. Thus, three unit standard scores are determined for each student: z_Q, z_T and z_E. Students are given rather complete descriptive numerical information for the current class after each quiz and the midterm test – the three most recent classes are included for the test.

To give the reader a feel for the variability of student performance, raw score and standard score distributions for the first course are shown as boxplots in figure 1. In general, over the years, student Q, T, and E scores have been fairly consistent – the three Pearson correlations for the past three classes (a total of 73 students) for the first course were in the 0.66–0.62 range.

The three standard scores are combined to yield a composite score:

$$Y = 0.5z_Q + 1.0z_T + 1.5z_E$$

For a number of years, students were asked to suggest weights used to obtain a value of Y. Virtually all classes arrived at, and agreed with, the above weights.

Each of the journal article critiques is 'scored' with a maximum score of 10. Similarly, a maximum score of 20 is obtainable for the research project.

Grading

Earned grades for each course are basically

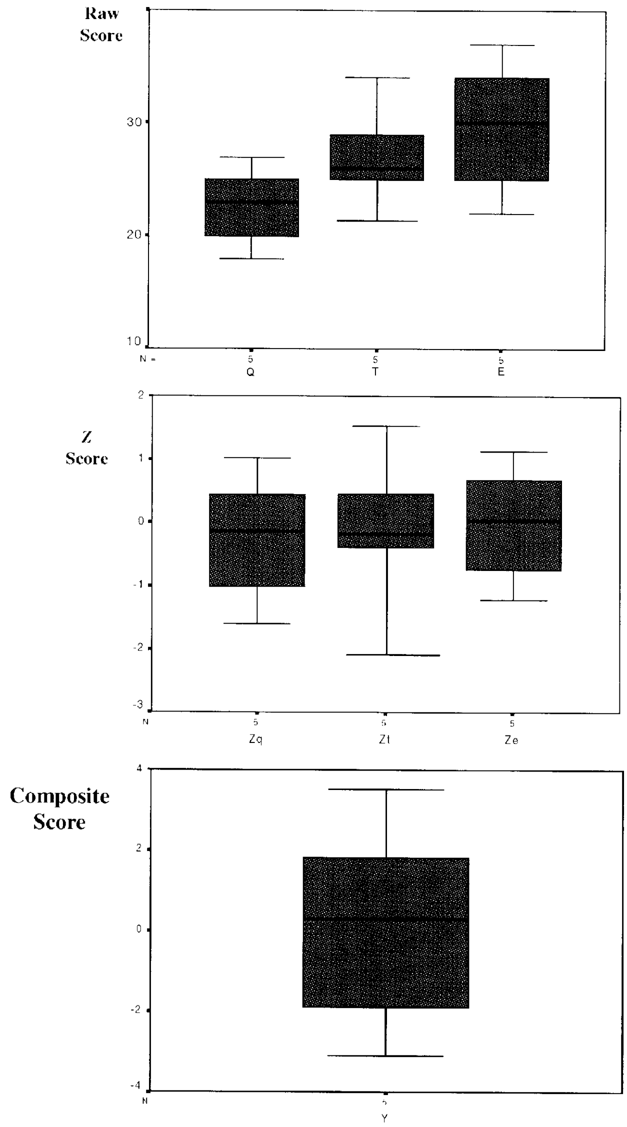

Fig 1. Boxplots for the most recent first-course class.

determined by the Y score. The following cut-offs have been used:

$2.00 < Y$	A
$-1.50 < Y < 1.50$	B
$Y < -2.00$	C

Even though the critique and project scores are not explicitly considered in the above, they are taken into consideration for students whose Y score falls in one of the two gaps. Of course, some judgement needs to be made. There are three reasons why critique and project scores are not included in the composite measure: (a) very limited score variability results; (b) the work involved is not that conducive to clear awarding of points; and (c) when two students collaborate, each is given the same number of points. The distribution of Y scores for the first course in the sequence is shown as a boxplot in figure 1.

This scoring and grading procedure has yielded somewhat varied grade distributions over the three-course sequence taught by the author over the past 25 years or so. These are given by the barcharts in figure 2. It is obvious, and expected, that the per cent of A grades increases over the three course levels, while the per cent of C grades decreases. Such trends may be expected over the three courses because the lower-performing students weed themselves out of taking a subsequent course.

◆ DISCUSSION ◆

Student assessment is not simple! It involves considerable effort on the part of the instructor, who is required to exercise judgements of various kinds at various times. But student assessment is very important, for both the student and the instructor. Generally speaking, student assessment should reflect some variability in acquired student knowledge and skills within classes, at both the undergraduate and graduate levels of education. It is recognized that there will be some classes (for advanced-level courses) in which there will be limited variability.

Occasional assessment throughout an academic term has been found to encourage student learning and has been very helpful to me as an instructor. Students are encouraged by the assessment to gain an understanding of concepts, and to appreciate some misunderstandings of concepts. With my instructional approach, most of the concepts covered by the assessments are re-emphasized in two ways after the assessments are taken: (a) in class discussions, and (b) in learning-group discussions. I have regularly learned whether some concepts need to be discussed again, instructional emphases redirected, or instructional pace changed.

How student performance is assessed to reflect expected student variability may vary across areas of study. An example of how graduate student performance is assessed in the area of statistical methods has been described above. No claim is made that this approach is applicable in all statistics courses. A certain degree of satisfaction has been achieved; improvements, however, are continually being considered and implemented.

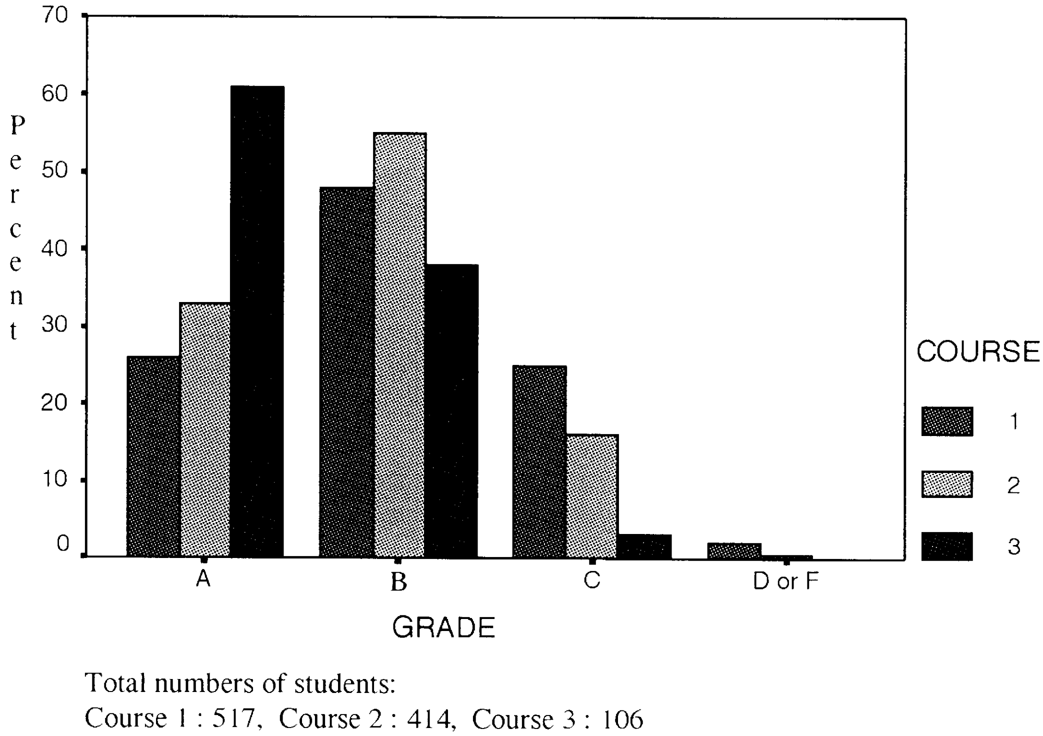

Fig 2. Grade distribution for the three courses.

References

Astin, A.W. (1993). *Assessment for Excellence*. Phoenix, AZ: Oryx Press.

Braskamp, L.A. and Ory, J.C. (1994). *Assessing Faculty Work*. San Francisco, CA: Jossey-Bass.

Brookfield, S.D. (1990). *The Skillful Teacher*. San Francisco, CA: Jossey-Bass.

Centra, J.A. (1993). *Reflective Faculty Evaluation*. San Francisco: Jossey-Bass.

Cobb, G.W. (1993). Reconsidering statistics education: a National Science Foundation conference. *Journal of Statistics Education*, **1**(1).

Eble, K.E. (1979). *The Craft of Teaching*. San Francisco, CA: Jossey-Bass.

Eble, K.E. (1983). *The Aims of College Teaching*. San Francisco, CA: Jossey-Bass.

Ericksen, S.C. (1985). *The Essence of Good Teaching*. San Francisco, CA: Jossey-Bass.

Gal, I. and Garfield, J.B. (eds) (1997). *The Assessment Challenge in Statistics Education*. The Netherlands: IOS Press.

Garfield, J.B. (1993). An authentic assessment of students' statistical knowledge. In: N. Webb (ed.), *National Council of Teachers of Mathematics 1993 Yearbook: Assessment in the Mathematics Classroom*, pp. 187–96. Reston, VA: NCTM.

Garfield, J.B. (1994a). Assessing learning in college statistics classes. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Garfield, J.B. (1994b). Beyond testing and grading: using assessment to improve student learning. *Journal of Statistics Education*, **2**(1).

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, **63**, 25–34.

Gullette, M.M. (Ed.) (1984). *The Art and Craft of Teaching*. Cambridge, MA: Harvard University Press.

Hawkins, A., Jolliffe, F. and Glickman, L. (1992). *Teaching Statistical Concepts*. Harlow, Essex: Longman.

Hubbard, R. (1997). Assessment and the process of learning statistics. *Journal of Statistics Education*, **5**(1).

Jolliffe, F. (1991). Assessment of the understanding of statistical concepts. In: D. Vere-Jones (ed.), *Proceedings of the Third International Conference on Teaching Statistics*, vol. 1, pp. 461–6. Otago: Otago University Press.

McKeachie, W.J. (1994). *Teaching Tips*. Lexington, MA: Heath.

Moore, D.S. (1995). The craft of teaching. *Focus*, **15**, 5–8.

Simpson, J.M. (1995). Teaching statistics to non-specialists. *Statistics in Medicine*, **14**, 199–208.