From **Gal, I. & Garfield, J. B. (editors).** **The Assessment Challenge in Statistics Education**. IOS Press, 1997 (on behalf of the ISI). Pages 191-204. ISBN 90 5199 333 1. Copyright holder: International Statistical Institute. Permission granted by ISI to the IASE to make the chapters from this book freely available on the Internet. The References section is a separate chapter and is available on the Internet at http://www.stat.auckland.ac.nz/~iase/publications/assessbkref.

15. Issues in Constructing Assessment Instruments for the Classroom

Flavia Jolliffe

Purpose

When constructing assessment instruments both the *purpose* of the assessment (feedback, grading) and the skills which are being assessed need to be considered. The main purpose of this chapter is to help teachers develop their own assessment instruments by giving specific examples of tasks. Unsatisfactory tasks are used to illustrate the pitfalls, and alternative versions are given as examples of good practice. Some comments on grading are included. The emphasis is on written assessment in the classroom, mainly of pupils aged about 14-19, but much is relevant also to introductory statistics courses at college and university level.

Consideration is given to ways of assessing factual knowledge, the ability to use computers, understanding of concepts and application of techniques, and communication skills. The pros and cons of multiple choice and open-ended questions are discussed as are the challenges of oral assessment and assessment of group work.

INTRODUCTION

In assessing student learning of statistics we need to assess factual knowledge, understanding of concepts, computational ability, appropriate application of techniques, and the practical skills of doing and communicating statistics. Although there are strong arguments for using practical work and projects as the principal methods of assessment of many of these qualities, other methods can be used successfully for most of them. Traditionally assessment has involved the student in answering written questions, possibly under a time constraint, but more recently some consideration has been given to the use of oral assessment.

Assessment can be formative or summative, a division corresponding approximately to the two main purposes of assessment, feedback both to the teacher and to the student (formative), and grading of the student (summative). In internal assessment in a classroom situation, which is the main concern of this chapter, feedback is more important than grading. All the points made, however, are also pertinent to external assessment by public examinations where grading is the only purpose.

Closed book timed examinations where students are required to do questions of a standard type have little to commend them as ways to assess student learning of statistics. Questions tend to be artificial in nature and are more likely to test short-term than long-term memory, and the time constraint may be unfair to some students. However, we can be fairly confident that the grading produced reflects the students' own performance. The provision of formula sheets reduces the reliance on memory and the extension to open-book examinations brings a new dimension into assessment and mimics the situation in the workplace. Some public examinations include analysis and discussion of real data (say one page of data) sometimes available to students before the examination. In other public examinations questions relate to newspaper articles (see also Chapter 10) or are based on reports of experimental investigations.

It is easier to see what is unsatisfactory with an assessment task than to start from scratch and write a good question, worded unambiguously in everyday language and meeting required objectives. Questions *can* be made crystal clear by using mathematical language of course, and we must use correct and precise terminology in teaching and assessment, but real life statistics questions are not posed in this way. Even when questions are subject to rigourous refereeing when being developed, students sometimes make a legitimate but different interpretation from that intended, or find unanticipated methods of solution. Bentz and Borovcnik (1988) consider explanations for alternative answers to relatively simple test items on probability. The issue must be more complex with more searching questions at a higher level.

This chapter starts the process of writing assessment tasks by giving examples of tasks and pointing out why some are poor and others are better. A general discussion of assessment issues is given in Chapter 10 of Hawkins et al. (1992). A framework for the generalisation of assessment tasks is given by Nitko and Lane (1991).

The design of assessment tasks should, as mentioned earlier, be guided by clear curricular objectives. For the purpose of this chapter, the following four goals were adapted from Hillyer (1979). These goals are compatible with and comprise part of the general framework of common goals presented in specific chapters in this volume.

1. To develop knowledge and an appreciation of the ideas involved when looking at information.

2. To develop knowledge and an understanding of methods of presenting information by charts and diagrams including those used in business and newspapers.

3. To develop necessary interpretive skills so as not to be misled by statistical information.

4. To develop a critical approach to the methods of collecting and presenting information and to the conclusions drawn from statistical data.

Below, I first discuss how to assess factual knowledge, then I discuss issues of assessment involving computers. After presenting aspects of assessing using multiple-choice and open-ended questions, I focus on the assessment of group work.

ASSESSMENT OF FACTUAL KNOWLEDGE

How can we assess factual knowledge? Why might we wish to do so? A fairly popular form of assessment question has been to ask for definitions of terms or formulae, particularly in class tests after pupils have been asked to learn material or in formal examinations; for example, "Define the median," "Give a formula for the arithmetic mean of a frequency distribution." Such questions test little more than the ability to retain material in memory and recall it correctly. There is not much point in this because in the workplace frequently used facts are learnt through their use and others are looked up when needed, but it *is* important to assess whether students know what to use and what to look up.

Let's start by considering testing factual knowledge of the median. We could ask for an interpretation; for example,

> The median number of hours spent doing homework by the 30 members of a class last week was five and a quarter. Write a sentence explaining what this tells us about hours spent on homework by class members.

Or we could ask when and why a median would be a suitable summary measure. Or we could ask a more specific question such as:

> In negotiating for a salary raise the union representative quotes the median income. The employer's representative quotes the average (arithmetic mean) income and this is higher than the median. Explain why these two summary measures can differ. Which measure do you think is more appropriate here, and why?

Rather than ask students to quote a formula it would be better to ask for a calculation, as they would need to know the meaning behind a formula in order to do this. For example, let's consider the arithmetic mean. The question below does not discriminate against those who have other ways of remembering how to find the mean than via an algebraic expression, such as those who use a mental picture of the layout of calculations. A formula sheet is not necessarily useful to such students.

> The following frequency distribution gives the number of days it took for a person to drive a certain distance. Estimate its mean to the nearest 0.1. Show any necessary work.

No. days	Frequency
0- 2	14
3- 5	29
6- 8	19
9-12	10

(Bernklau)

Another possibility is to ask a question which tests understanding of a formula or concept as in the examples below.

> Circle the *single* best answer.

Leonard made an 84 in the first test and an 89 in the second test. If all tests are weighted the same and Leonard's teacher does not round averages, what does Leonard need to make on the third test to have an average of 90?

a) 90 b)96 c) 97 d) 98 e) 100

(Kaigh)

A set of scores has a mean of 68.7. If 10 were added to each score then the mean would be ......

(Bernklau)

There are good pedagogic reasons why students should have some familiarity with formulae and should do some simple hand computations. Both aid understanding and help remove the black box aspects of calculators and computers (Jolliffe 1991, 1993), but they should be seen to be short first steps and should not be laboured to such an extent that they become stumbling blocks. The principles behind methods of making computations easier, such as change of origin and scale in finding means and variances, are important but should be presented briefly.

Another way of testing factual knowledge is to ask for explanations of how to perform operations or the meaning of terms; for example, "Explain how to use a graphical technique to estimate a median from a grouped frequency distribution" or "Explain what is meant by simple random sampling." One problem here is that those who have memorised notes or portions of textbooks may be able to reproduce appropriate sections without understanding (and some might write out the wrong section!). Another is that giving explanations in the abstract is quite difficult. Thus a question asking for an explanation might be improved if it also asked for an illustration. For example:

Explain how to use a graphical technique to estimate a median from a grouped frequency distribution. Illustrate your answer by estimating the median of the frequency distribution below which ....

Explain what is meant by simple random sampling. Describe how you would select a simple random sample from the list of 200 people attending a one-day meeting.

An alternative to providing an illustrative example within the question is to ask students to describe examples with which they have been involved. This has the advantage of linking the assessment to practical statistics. But if students have access to the material they could just copy out examples; if they do not the question is mainly one of memory.

COMPUTATIONAL ABILITY

Computational ability is important and students need to practise questions on computation, but teachers should not set large numbers of such questions just to keep a class occupied, nor base a large proportion of assessment on such questions. Questions which do little more than ask for a computation— for example, "Calculate the standard deviation" or "Fit a regression line," —test whether students can choose the appropriate function, and do arithmetic or use technological aids correctly, but do not test ability to apply principles in statistics. Both factual knowledge and computational ability can and should be tested through questions which are of statistical interest.

Students should be shown how to obtain approximate answers with pen and paper so that they develop a feel for the magnitude of answers and can recognize when an answer is clearly wrong, perhaps because a wrong function button has been pressed on a calculator. Ways of testing whether students are able to do this include: asking for an approximate answer, presenting the question in multiple choice form, asking for an explanation as to why a given answer cannot be correct, and asking for eye-ball comparisons of data sets.

Students who get an impossible answer at any stage such as a negative variance, a probability outside the range [0,1], or a correlation coefficient outside the range [-1,1] or where the sign is clearly wrong, should be penalised severely if they do not comment that it must be wrong because this shows lack of factual knowledge and of understanding.

ASSESSMENT OF ABILITY TO USE COMPUTERS

As statistics packages (including spreadsheets) and word processor packages become more readily available in or out of the statistics classroom, students will be making increasing use of a computer for statistical work. Assessment of their ability to do so will therefore take on a greater importance. The type of assessment tasks need not differ fundamentally from those in which the use of the computer is not required, and all the principles of good and bad assessment apply here too. Many papers have been published describing experiences of using the computer in teaching and advocating its use, but few mention assessment. An exception is the paper by Phillips and Jones (1991) in which they state briefly how they aim to make assessment reflect the style and flavour of a statistics course for engineering students in which computing is an important component.

The skills we might wish to assess include knowing what analysis it is appropriate to do, the ability to get a package to perform this, and interpretation and editing of output. Assessment of knowledge of what statistical analysis is required is almost no different when computers are used as when analysis is done by other means, although the choice is clearly limited by what packages are available. This is likely to become less of a problem in the future as simpler packages become more sophisticated and computers become more powerful and user-friendly.

Assessment of practical skills such as importing data and other files, and being able to obtain a hard copy of output, possibly in edited form, might be assessed in computer science courses and study of methods used there could be helpful. Skills such as being able to access and use a package, including data entry, are more specific to statistics. Actual computing skills as used in spreadsheet manipulations or in writing routines to use within a package could possibly be assessed in the same way as a “hand” computation. Assessment of the selection of interesting and appropriate parts of the output and interpreting them should present no particular problems in that the role of the computer is that of a tool like a calculator.

The teacher needs to avoid the temptation to overwhelm students with data. Just because packages can deal easily with large data sets does not necessarily mean that it is good for students to spend a lot of time looking at large data sets. Students do not have to do all the computing themselves—they can be asked how they would use a package and to say what commands they would choose, and they can be asked to interpret a pre-prepared piece of output. The following is an adapted example of a question which included computer output showing commands, dotplots and stem and leaf diagrams.

Fuel consumption figures for 57 types of car are reported in miles per US gallon in the following Minitab output. The data come from issues of the US *Consumer Reports* magazine between 1987 and 1989. The data show fuel consumption for city driving, an open road trip and expressway driving on an interstate highway at a steady 55 miles per hour.

(a) From the data shown, what is the most economical fuel consumption achieved with expressway driving? What is the median fuel consumption for open road trips?

(b) Write a brief report (no more than 200 words) summarising the conclusions you draw from these data about fuel consumption, covering both the way it varies between different types of cars and the way it is affected by the type of driving done.

(Fuller)

Use of computers in an exploratory way, for example, building up sampling distributions or seeing the effect of outliers on regression lines, would not normally be graded (but see Chapter 19). Computer aided learning and assessment is a subject in its own right and is not discussed here, but it is worth noting that feedback is provided when wrong answers are given and that records of scores obtained can be made available to the teacher, as noted by Lajoie in Chapter 14.

MULTIPLE-CHOICE QUESTIONS

Multiple-choice questions have become fairly popular in recent years. They are quick to administer and mark/grade, but on the whole are suitable only for short questions requiring a minimum of thought. They cannot probe deeply into understanding or require long answers, though they sometimes ask for a justification of the response. Multiple-choice tests are as much an aide memoire to the student as useful to the teacher and are perhaps most suitable for formative assessment or for quick feedback.

A serious objection to multiple choice is that the distractors might reinforce incorrect ideas. Giving three choices is fairly popular. The student then has a 1 in 3 chance of guessing the correct answer which should be varied in position. A simple right/wrong scoring system might be used. Allowing students to select a “Don’t know” option should help reduce guessing and the instruction “Give the MOST CORRECT answer” helps to convey the idea that there is not always a unique correct answer in statistics.

Many multiple-choice questions are bad because there is no attempt to relate them to a context, and it is not clear what is being tested other than an ability to apply a definition and perform a computation correctly. An example is:

Given the probabilities of x = 1, 2, 3, 4, 5 being 0.1, 0.4, 0.2, 0.2 and 0.1 respectively:

What is the expected value of x-squared?

a) 5.0 b) 7.84 c) 9.2 d) 6.25 e) 3.03

What is the standard deviation of x?

a) 1.166 b) 1,360 c) 2.8 d) 1.0 e) 1.44

The following questions are better because they test concepts and do not require much computation, but these too can be criticised for their lack of context and realism. An improvement might be to relate them to scaling of examination marks/scores.

> The mean of the numbers 5,6, and 7 is 6. If each number is squared the mean becomes:
> a) 36 b) greater than 36 c) less than 36
>
> The standard deviation of the numbers 5,6, and 7 is 1.
> If 1 is added to each number the standard deviation becomes:
> a) 1 b) 2 c) square root of 2

Care has to be taken in setting questions of this type that the answer is not so obvious that the student hardly has to reason in order to get it correct. On the other hand it should be possible to select the answer without doing elaborate calculations. This is not as easy to achieve as it might at first seem. Studies have shown that people are rather poor at "subjective" estimation (Hawkins et al., 1992, Chapter 6).

Multiple-choice questions are used in research studies, for example, to find out misconceptions in order to inform remedial action, or to assess the success of an intervention such as practice simulation exercises. Such questions are not designed for assessment purposes and adverse criticisms of multiple-choice questions do not necessarily apply to them. However, some of these are good models for assessment questions, particularly some of those used in studies relating to probabilistic and statistical concepts. In research it is common to follow up the question with "Why do you say this?" or equivalent.

The following example which is taken from a statistical thinking survey (Swets et al., 1987) is a good example of the kinds of questions asked. The instructions included the statements:

> Please answer each question as well as you can, paying particular attention to the 'Why' part. If you have an idea you are not sure is correct, include it as well. We are especially interested in how you think about these problems.
>
> > *If you wanted to investigate whether the Beverly Hills High School students were better or worse in mathematics than the Valley High School students, check which of the following methods you would choose:*
> >
> > () Pick one student at random from each school, give them the same math test, and compare their scores;
> > () Give the same test to every student in the two schools, and compare the total of the scores obtained in each school;
> > () Give the same test to every student in the two schools, and compare the highest score obtained in each school;
> > () Give the same test to every student in the two schools, and compare the average score obtained in each school;
> > Why would you use this method?

Chapter 16 (Wild, Triggs, & Pfannkuch) in this volume focuses in greater detail on multiple-choice questions, their limitations and benefits, and reasonable ways to use them in contexts where other formats are impractical, especially in assessing large college classes.

OPEN-ENDED QUESTIONS

Open-ended questions are a complete contrast to multiple-choice questions and to very specific questions which say what is to be worked out and which lead students into appropriate comments. Probably the most open-ended question which can be devised is of the type “Analyse the following data and report on your findings.” The problem with such questions is that as there is no guidance as to what is expected; attempts at them are likely to vary greatly both in quality and in quantity. Students may do every possible analysis imaginable, including some which are inappropriate. It could be argued that questions of this type are mini-projects and could be assessed as such. The assessor needs to decide what analysis it is reasonable to expect the student to do in the time available, and to have a grading system or scoring rubric (see Chapter 3 by Colvin & Vos) which is flexible enough to allow for non-standard and unexpected approaches.

Short open-ended questions involving an element of reasoning can be useful. For example:

> What, if anything, is misleading about the following claim?
>
> *An increase in sample size will always cause a sample mean to move closer to a true mean.*
>
> (Bernklau)

A slightly longer task involving real data is:

> An *El Paso Herald-Post* article discussing sports-related deaths stated “Basketball is the country’s most lethal sport.” This “rather startling finding” was based on 1989-1991 data from the National Center for the Study of Sudden Death in Athletes, which reported the following death numbers: basketball, 70; football, 42; track, 38; baseball, 24; skiing, 16; soccer, 13; weightlifting and hockey, 10 each; wrestling and boxing, 8 apiece; golf, 7. Comment on the conclusions quoted from the article.
>
> (Kaigh)

A more focused task involving analysis of data will ask a specific question and perhaps hint at what is required in the way the information is presented. The following is a good example of how to do this.

> A study was conducted on highly intelligent children. The researcher collected data on 30 children on age and IQ. The data are below. The researcher claimed that one student, Dean, scored higher than most students on IQ, but not on age. Is this claim acceptable?

Student	**Age**	**z score**	**IQ**	**z score**
	(years)	(Age)	(IQ)	
Ed	12.10	1.07636	143.00	-.86936
Sam	9.40	-1.13647	152.00	1.82864

Dean	13.00	1.81398	148.00	.62953
Ann	9.30	-1.21842	140.00	-1.76869

(note: the original table is much longer than the portion shown here).

(Cohen)

We might well wonder how students who answered “Yes” or “No” without further comment would have been graded! A better wording might be “Perform an appropriate analysis to help determine if he can substantiate his claim.”

A question might leave the decision of fine detail to the student but give some guidance as to what is needed. For example, a question which says “Perform an appropriate test to investigate whether group A scored higher on average than group B,” where scores are given for members of both groups, is open-ended in that the student is left to choose the test, state the statistical hypotheses, and so on. A question saying “Make two brief comments comparing the given distributions” gives a little guidance as to the kind of response expected and its length. One with a wording such as “Summarise the data in a manner suitable for inclusion in a company report” has an element of open-endedness in it and a touch of realism.

Graham (1990) suggests how to find suitable statistical investigations and how to modify closed tasks to make them more appropriate to the PCAI cycle (Pose the question, Collect the data, Analyse the data, Interpret the results). For example, instead of the closed task “Find and record the number of pupils born in each month of the year,” it would be better to ask “Is it true that more babies are born in the Autumn than other seasons?”

Many texts contain case studies, some of which are in the form of open-ended problems. Morris (1989) builds up a case study, presented as letters from Jane, who runs a catering business from home, throughout the book. One letter asks for an explanation of MINITAB output of a customer survey, and the instruction is to write a short report interpreting the information in practical terms, making recommendations for action. In another letter Jane is considering whether to build an extension to her kitchen or to buy or rent other premises and asks how decision theory works. The exercise is to draft a reply using the basic situation and inventing data to illustrate the ideas of decision theory.

These examples show how open-ended questions can go a long way towards meeting the requirements of good assessment questions. They are able to pose “real” questions with “real” data and assess choice of appropriate techniques and the doing and (written) communicating of statistics. They also ask students to explain concepts—a sure test of understanding. However, if there is an obvious link between such a question and the subject matter which has just been covered, the question will not test the student’s ability to tackle an unknown problem or to bring together a range of methods learnt throughout a course.

An essay question is another type of open-ended assignment, particularly useful for indirectly exploring understanding. We might ask the student to develop or evaluate a research design, or write about the evolution of statistics, or the use of statistics by the media. A quantitative essay with an emphasis on numbers rather than words is another possibility and might usefully link with the student’s other subjects. For topics such as “Is defense a major item of government expenditure?” or “The measurement of poverty,” the student would need to find relevant data, present this as tables and/or diagrams, discuss the facts using summary or other measures, and suggest further analysis where appropriate. Giving students examples of papers or books where the mix of words and statistical aspects is about right, such as articles in *Chance* (American Statistical Association), will guide students as to what is expected.

Scoring

Almost certainly the grade awarded to an essay will be based on a general impression—something that arts teachers are used to doing, but a little alien to the typical statistics teacher. However, essays are a form of assessment well worth considering. They test the ability to organise material from a variety of sources, may test creativity, and almost certainly test understanding. They can be vocational, they allow students to draw on and demonstrate their own knowledge of statistics, and on the whole are not memory dependent.

Regarding scoring of open-ended questions in general, a cross between subjective and objective grading is one way of assessing attempts at open-ended questions. The assessment might be broken down into putting the problem in context, appropriateness of analysis, computational accuracy, and conclusions with each aspect graded on a 5-point scale. Grades could be combined if required. A scoring method developed by Garfield (1993) uses a 0/1/2 scoring with 0 indicating incorrect use, 1 indicating partially correct use, and 2 indicating correct use. The categories she suggests are communication, visual representations, statistical calculations, decision making, interpretation of results, and drawing conclusions. This system adapts well for use in other contexts.

ASSESSMENT OF GROUP WORK

It is important to train students to work in groups because people doing statistical projects need to be able to interact with others; this is a useful life skill in many fields of employment. Sometimes sharing of effort may be the only practical way of completing a problem in the time available. However, the assessment of group work poses many challenges. In general, it involves assessing two separate but related issues: the assessment of the nature and quality of the group *process*, and the assessment of the quality of the *product* of the work of the group. In both cases, it may sometimes be useful to be able to determine the individual contribution of each student to the process or the product, and consider whether each member is given the same grade (i.e., the group grade), or whether adjustments are made depending on the difficulty level of the subtask a student handled, his/her initial skill or knowledge level, and so forth.

As for assessment of process, this topic is handled in some detail by Curcio and Artzt in Chapter 10 of this volume. The product of the group work could be judged holistically, with a score given on a scoring rubric with anywhere between 3-6 levels. Or, separate categories could be established for each aspect or facet of the group product (e.g., problem definition, design/approach, methods of analysis, interpretation, conclusion, presentation), and a weighted score given to each aspect. Often, a way of assessing group work is to award the group a single overall mark based on marks/grades given to the various components of the task performed by the group (see also the scoring rubrics used by Keeler, Chapter 13; or Colvin & Vos, Chapter 3; in this volume). The assumption here is that whether or not every group member has been involved to the same extent with the execution of every aspect of the work, all should have been involved to some extent, for example, in planning and in deciding who would do what. An example of a scoring rubric that can be used in such assessment can be found in Garfield (1993).

Some teachers may want to give a group project a single composite score that takes into account both the quality of the product and some features of the process (e.g., whether the group collaborated well and made effective use of all members' skills and contributions). This is possible but will depend on the amount of information available to the teacher about the group process. In some contexts teachers may be helped by using peer assessment, as discussed below in Implications. One way of assessing the contribution of individual students to a group process involves using a fairly broad grading such as a 5-point scale with a weighting system to take account of the difficulty of the task and (mis)matching between students and tasks. A task that was inherently hard might merit increase of one grade and a task that was too hard for a student might also merit increase of one grade, with corresponding decreases for inherently easy tasks or tasks too easy for a student. There is of course subjectivity in this. Even when we think we are making an objective assessment there is usually an element of subjectivity in there as well.

When students are all doing the same thing but pooling results, each student can be assessed as an individual. If in group discussion about the task there is variation in the quality and quantity of inputs of different students, a mark might be given to reflect this (provided the teacher feels able to make a fair discrimination—not necessarily an easy task!). To some extent a similar problem occurs in any assessment in that some students may be leaders and others followers or sometimes mere copiers, and it is not always possible to detect where credit for originality is due.

In a project type investigation students might work on different aspects; for example, some might concentrate on background reading, some on data collection, others on analysis or report writing. Here assessment poses a real challenge. This is because:

- Tasks may differ in the nature of what is required or the level of difficulty or the time needed to complete them.

- Some students might have tasks on which they are able to perform well, whereas others have tasks they find hard and on which they are not able to demonstrate their highest level of attainment. For example, a student who excels at data analysis but is poor at report writing is likely to get a higher grade if doing analysis than if writing a report.

- The method of assessment appropriate to one task might be inappropriate for another; for example, a fairly "accurate" scoring system might be used for data analysis but a subjective grading system might be better for a report.

- Different stages of the work do not always lead to the same type of output; for example, thinking and planning might lead to no tangible hard copy.

- It may be difficult to sort out individual work from group effort and to assess interaction among group members—an important aspect which should not be neglected in assessment.

In fact it could be argued that awarding individual marks to group members is contrary to the spirit of group work, but sometimes it is important to discriminate between students, for example for the award of scholarships. One way of modifying a group's mark to take account of an individual's contribution is to ask students to rate each other (and even themselves) and devise a weighting for each student based on an average of the ratings (s)he has received including perhaps a rating from the teacher. Again it is difficult to avoid subjectivity but peer assessment is in itself another workplace skill and it does no harm to expose students to it early on. Another option is to use an observation form designed along the lines suggested by Curcio and Artzt in this volume (Chapter 10).

ORAL ASSESSMENT

The relatively new development of oral assessment in statistics is to be applauded because oral presentation of results is as important as written presentation and is something that occurs in the workplace, but there is as yet little experience of it. As with any form of assessment, standardisation of questions and techniques is important, but the administration of oral assessment is more obviously uncontrolled than written assessment, and ideally teachers need some practice in it before using it for summative assessment. A partly subjective broad grading scheme may work fairly well, most likely using a rubric with 4-5 levels, or several categories with an attached weighting scheme.

Oral assessment can be part of a more traditional assessment; for example, students might be asked to clarify written answers or to explain their methods. On occasion it might substitute for other methods. It is useful for assessing young children as when the teacher reads pre-prepared questions and asks students to indicate correct answers on a prompt sheet, for example, to point to the modal group in a histogram. Oral assessment can also be useful for diagnostic purposes to identify points which have been misunderstood or problems experienced by low achievers.

Tape recording or making a video of an oral procedure means that grading can be done independently of the administration of the assessment but could have an inhibiting effect on the students (and perhaps the teacher). Smith and Griffin (1991) describe the use of videos in training and Chapter 14 is also relevant. Negative aspects of oral assessment are that those whose spoken language facility is poor, typically those who are being assessed in other than their mother tongue, those with speech difficulties, and very shy students could be disadvantaged.

IMPLICATIONS

In sum, the key points of this chapter are as follows:

- Setting good assessment tasks can be tricky, but there are many good examples on which to build. Always remember—if you come across a good assessment question keep it carefully!

- Factual knowledge is best assessed through questions where the main aim is to test understanding. Such questions might involve some computation.

- Computational ability can be tested as part of a more challenging problem. We should teach students how to obtain approximate answers but should not necessarily assess them on this.

- Assessment of computer use involves statistical and computing aspects. Assessment of the former is similar to any statistical assessment, of the latter as of any computing tasks, but this is a new area and there are not yet many published examples of it.

- Multiple-choice questions are sometimes useful, but take care not to overdo their use.

- Open-ended questions can be long or short. They can give some guidance as to what the student is expected to do. They are ideal for teaching statistical understanding and skills but grading is partly subjective which the statistics teacher might find a little strange.

- Group work should be encouraged but assessment of it presents new challenges. Little experience exists in this area and there is room for exploration and experimentation.

Many of the traditional methods of assessment in statistics concentrated on rote learning and computational aspects. They are now seen to be unsatisfactory in the new climate of teaching and learning statistics where the understanding of statistical concepts is emphasised. Open-ended tasks, preferably using real data, can be made an important part of the teaching process and are valuable learning tools, but marking and grading students’ attempts at such tasks is a new area for many teachers. Multiple-choice questions and questions where the method is prescribed are less satisfactory in assessment, but used with care and in moderation can be useful—and grading does tend to be easy.

The increasing use of computers, and of group and oral work in statistics is to be applauded, but their use in assessment is a relatively unexplored area, and the development of suitable assessment tools with associated guides to grading is required. The practising teacher has an important role in contributing to the growing body of knowledge about these methods.

It is an exciting time as regards methods of assessment. Don’t be afraid to experiment in your use and development of assessment tasks and grading methods—be imaginative and innovative. Use a variety of methods for a richer insight into student learning. Above all be critical of your own and others’ assessment tasks. Note what works well and what does not and share your findings.

Most of this chapter has been concerned with assessment by the teacher, but the students themselves could mark their own or others’ attempts at some of the more routine assessment tasks if provided with suitable “model” answers. In the case of formative assessment this could be quite satisfactory. Making random checks on the process would help ensure students were honest.

Part of the purpose of assessment is to monitor progress and to take appropriate action if students are not performing satisfactorily. An interesting possibility is to get students to use quality control techniques to monitor their own progress. As higher and lower control limits each student could take the highest and lowest scores (s)he expects to obtain on assigned work. More details and a pro forma are given in Kimmel (1992).

The context in which statistics is presented is important. If students do not care for the context they will not be motivated to learn and in consequence will perform badly on assessments. We need to be careful that we do not inadvertently form a wrong opinion of the statistical expertise because the student has been turned off by the context; for example, research has shown that females perform better on “people” questions than on male oriented or abstract questions (e.g., Clark, 1994). Many recent texts use real data or base questions on actual studies. This is likely to motivate students but the teacher may need to be selective as regards the emphasis in the real data taken from texts and other sources for teaching and assessment purposes. In particular these should be gender-free and should not be biased for or against any particular ethnic or social group.

These ideas are complemented by other chapters in this volume, especially those discussing “alternative” assessment, such as Colvin and Vos, Keeler, and Lesh. But more research needs to be done, both in traditional “academic” form (e.g., about assessment with computers), and more importantly, in the form of “teacher research” or “local research.” In these latter forms teaching teams or departments interested in revamping their assessment schemes may try to selectively implement new forms of assessment and see what is the value of the new information obtained for informing teaching and giving feedback to students, or what is the impact on students, e.g., of using certain scoring rubrics, or giving some credit to group process, not only to group product.

Acknowledgments

Thanks to David Bernklau of Long Island University, Steve Cohen of Tufts University, Mike Fuller of the University of Kent at Canterbury (UK), and W. D. Kaigh of the University of Texas at El Paso for their willingness to make their assessment questions available for inclusion in this chapter.