From **Gal, I. & Garfield, J. B. (editors). The Assessment Challenge in Statistics Education**. IOS Press, 1997 (on behalf of the ISI). Pages 205-220. 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.

16. Assessment on a Budget: Using Traditional Methods Imaginatively

Chris Wild
Chris Triggs
Maxine Pfannkuch

Purpose

Because many of the authentic assessment methods described in this book tend to be very demanding of teacher-time, there is still an important place for finding ways to employ inexpensive, traditional methods more creatively in an attempt to come closer to achieving the same goals that those who advocate authentic assessment methods are targeting. The basic idea is to identify the elements of statistical thinking that we want to foster and then find ways of testing these elements with objective assessment methods (in particular, multiple choice). This chapter explores the extent to which objective testing can approximate the results of authentic assessment techniques and the extent to which it falls short. We provide guidelines for the writing of objective test items, together with examples.

INTRODUCTION

The authors are part of a team teaching introductory statistics to 2,600 students (200-300 per section) in a research university on a tight budget. Although the first-year statistics course that we teach was once primarily concerned with statistical calculations, we have now become more concerned with producing “intelligent citizens” than specialist statisticians. The majority of our students will not become statisticians, but all will be users or consumers of statistics in their careers and daily lives. We live in a world where, more and more, machines do calculations and data processing. The essential human role is to ask questions, and to obtain, interpret, and synthesise information from a variety of sources. We want to help our students to prepare themselves for this new world. We want to embed our teaching in the rich context of the investigative process, to foster the abilities to carry out investigations, to critique the investigations of others, and to communicate what data are saying. Thus, we are striving for the same general goals as those presented in the first part of this book. The most powerful signal that we have for telling students what we believe to be important is assessment.

In our university, there are pressures to improve teaching as well as increase the research output of faculty, but the main institutional focus is on the latter. With continued downward pressures on the funding of higher education across the English-speaking world, more and more teachers of statistics seem fated to have to do more with less as faculty/student ratios erode and classes grow ever larger. Despite the pressure for increased "efficiency," we are striving to pursue the same goals as those who advocate authentic assessment methods in statistical education. However, because many of the assessment methods in this book tend to be very demanding of teacher time, our challenge has been to describe our search for ways to employ inexpensive, traditional methods more creatively in an attempt to pursue the same ends.

The basic methods that we use—assignment work and multiple-choice testing—are probably an anathema to most contributors to this book and yet the time savings, when applied to 2,600 students, are overwhelming. The focus of this chapter is multiple-choice testing. Wild, Triggs & Pfannkuch (1994), an earlier revision of this chapter, contains a fuller discussion including the strategies we apply for assignment work to remedy the deficiencies of multiple choice. Although multiple choice cannot do at all well some of the things we value most, we have been pleasantly surprised by what we have been able to achieve and at some of the positive side effects on our teaching. This chapter explores some strategies for making the thinking required in a forced-choice test much more like that required in a real statistical investigation, and it explores the pros and cons of multiple-choice testing using a sample set of questions. Many of the suggested strategies apply just as much to free-response testing as they do to forced-choice testing.

This chapter is structured as follows. The following section, entitled "The Assessment Challenge in Large Classes," outlines the principles that we believe in with regard to assessment and the reasons that led us to adopt forced-choice testing techniques, and goes on to introduce the authors' techniques for overcoming some of the difficulties. This is followed by a section called "Practical Examples," which contains a suite of questions that illustrate the use of these techniques. All this becomes background to the deeper discussions in the following section—"What Have We Learned?"—which has subsections entitled: "Costs"; "Educational advantages"; "Considering some criticisms"; and "What multiple choice cannot do." The chapter closes with "Implications".

THE ASSESSMENT CHALLENGE IN LARGE CLASSES

Principles guiding assessment

The focus for assessment has traditionally been on certification, that is, as a measurement device to enable teachers to certify that a student has a certain level of mastery of ideas/material. More recently, the roles of assessment in motivating, focusing, and monitoring the learning process, and monitoring the interaction between teaching and learning, have been receiving more emphasis. Fortunately, these roles are complementary. Whether we like it or not, the assessment materials we give to present students and have given to past students strongly influence where students put their efforts. To a very large extent, what you test is what you get. Assessment that guides student attention in less than optimal directions fails in the certification role. Finding holes in student learning for certification purposes also exposes areas where teaching could be improved. Curricular goals should be formulated to meet the needs of the "customers" of the course, e.g., the students themselves, employers, society at large and teachers of future courses

(Wild, 1995). Assessment should endeavour to assess the extent to which these goals are being met. Because no form of assessment is perfect it is better to test the important inadequately than it is to test the unimportant well. In addition, assessment must be manageable within the budget (of people, money and time), and be perceived to be fair by students. General considerations about goals and assessment are considered in the first part of this book.

Constraints

Our first-year course runs through a full academic year. We have 2,600 students who are taught and assessed by 12 lecturers who should each be allotting less than 25% of their time to this course. In some years, as few as two of the lecturers are experienced as both statisticians and teachers. Most of the rest are Ph.D. students who are relatively inexperienced in both roles, although we usually have one or two who are experienced as teachers but not as statisticians. Backup tutorial assistance is provided predominantly by Masters students, and there is enough marking (grading) assistance (mainly undergraduate) for about 9 assignments, each of which take the students about 3 hours to complete.

Our switch to multiple-choice testing was entirely a resource decision, a way of saving time. Our assessment challenge is to get as close as we can to the educational goals we have outlined above, to emphasise statistical thinking and not just routine mechanical tasks using the tools at hand. But it has not all been bad; there have been some quite positive side effects that will be discussed later. Putting additional effort into assessment will necessarily result in something else suffering because almost all the people on the teaching team are already committing far more time to their teaching than is good for their careers or for the fulfilling of institutional priorities. We are not unique in this, for no one lives in a world of abundant resources. Everyone is in the business of trying to put a limited set of resources into the places where they will do most good, with all the uncertainty and under-informed tradeoffs that implies.

Our stratagems

We cannot do everything we want to do with multiple-choice methods and need other forms of assessment as well. Philosophies underlying the assignment work that complements our tests and examinations are discussed in Wild, Triggs & Pfannkuch (1994). However, we have found that we can overcome some of the inadequacies that are often claimed to be inherent to multiple-choice testing. We can do something about incorporating synthesis and evaluation, about minimising rote learning as a primary study focus, about testing big ideas rather than merely subtleties, and about interrelationships between ideas. We can also do something about simulating the stages of an actual analysis of data.

We use the following stratagems to accomplish these tasks:

1. Begin by collecting a file of real stories/data sets.

We do not think about types of questions at this stage except in the broadest of terms. Ideally, these data sets are so sufficiently rich in context and features that many questions can be built around a single story. Although we have only presented 8 questions about the data set in the next section, we could easily have asked 30. We would typically have five or six data sets in a threehour examination, although not all of them would be as elaborate as the one here. (We have never found it necessary to use artificial "data" and would view doing so as failure.)

2. Present the background, data, and a fairly large array of numerical and graphical summaries derived from it as a complete package.

All the information is presented together in preference to releasing the information in small pieces with the questions as they require it. This strategy forces students to synthesise and make use of a body of information. We believe in supplying more information than will be used so that students have to become selective. Most of the questions in the "Practical Examples" section (to follow) come from an actual examination.

3. Try to ask as many questions from the same story as possible, letting the story/data/situation suggest the questions.

Some reasons for this are as follows. First, we are more likely to get some real statistical thinking into the exam paper. Second, it is a way of triggering new ideas and fresh questions. Third, with only a few stories we can afford more complexity, and students can afford to make the time investment necessary to understand the context. Having fewer stories also shortens the effective paper length.

4. Include questions of the following five main types:

We have identified 5 aspects of the statistical process and statistical understanding that assessment should address. These are: (i) critiquing practical aspects of studies; (ii) interpreting data-information and making inferences from it; (iii) interpreting and understanding statistical ideas; (iv) specifying what techniques should be used in a given situation; and (v) performing mechanical tasks (e.g., calculations).

5. Break down tasks into subtasks and examine each subtask (alternatively, a coherent set of ideas) separately.

We do this in part because we cannot give partial credit, but there are more important reasons. This division of tasks reduces the gap between the answer to a question and the thinking processes used in its construction—the only information the teacher gets is the answer. If several skills are mixed up in a question, then if students get it wrong we do not know where the problem lies. If the question does not have a coherent theme, it will not be useful as a means of focusing the attention of future students on important points. Thoughtful question writing that tries to get at thinking processes can motivate us to isolate exactly what it is that we are trying to teach. Unless routine arithmetic is being tested, all necessary summary statistics (and a few plausible distractors) are given. Where answers are numerical, most alternative answers are obtained by using the types of mistakes students commonly make in written work. With verbal concepts, we try to use common misconceptions.

6. Provide other information to cut down on unproductive rote learning.

There are many ways of doing this, from conducting open-book examinations to allowing crib sheets. Our examination papers incorporate almost all of the formulae that students will need to use. We try to make conscious decisions about what is worth committing to memory and what is not. We make almost no demands upon students to remember formulae.

7. For questions about interpreting statistical concepts, or for "what to use where?" questions, bury one false statement amongst a collection of true statements.

To examine the understanding and interpretation of statistical concepts, and knowledge of when and where particular tools are appropriate and useful, we construct questions which ask students to identify the one false statement in a set of five where one statement is false and the other four statements are true. The true statements are usually important statistical messages, concisely stated, that we wish to transmit to future students. This reinforces the teaching and minimises the amount of random misinformation being fed into the brains of future students when they use old tests in their exam preparation. The correct answer can be obtained either by recognition or by eliminating the other possibilities. If students are being asked to critique something in particular, e.g., a particular plot, or anything to do with a particular data set, then one also has the option of asking students to identify the true statement among a set of false statements. Because this type of information is transient, it does not matter if much of the information presented is false. See Frith & MacIntosh (1984) for other ways of constructing multiple-choice tests out of true and false statements.

8. Avoid statistical jargon in interpretational questions.

The statements used in interpretation questions should be expressed in language that is as close as possible to plain English. Why? Because we believe that the cycle of real statistical investigation begins with questions that are intuitively understandable and nontechnical and has not been successfully concluded until conclusions are understood on this same level (Wild, 1994). An important component of the ability to communicate statistics is to be able to interpret the results of abstract manipulations in terms of concrete reality. It is all too easy to work successfully in an abstract framework with no real understanding of what it all means. Thus, we have to take care that the language used is no more complex than it needs to be. Statements about statistical ideas that are concise, correct, and easy to understand can be extremely difficult to write.

We believe that most of these eight recommendations apply much more generally than just to multiple-choice testing. We would want to use every stratagem discussed above (except for stratagem 5 and stratagem 7) for a time-limited free-response test. In a free-response examination, the intent of stratagem 3 (asking a multitude of questions) may be approached from the opposite direction by asking one or two totally open-ended questions and allowing students to dig around in the available information and reach whatever conclusions they can from it. Without signposts to stimulate thinking in the host of directions possible, one may get some very blank answer books because students have had little opportunity to display what they do know. However, the ability to handle such open-ended questions is obviously what we should be aiming for.

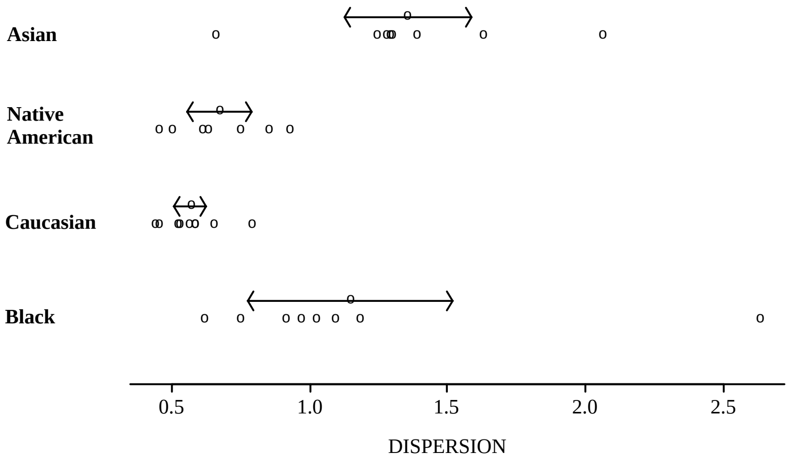
PRACTICAL EXAMPLES

Most of the questions which follow come from the last examination that the authors worked on together. They exemplify many of the strategies described above. The original examination paper also included traditional items requiring using numerical information that had been provided. These skills are still important and still tested, but no examples are presented here. The questions that follow are based upon research about measuring human genetic damage resulting from environmental exposures (Margolin, 1988). It comes from real research (stratagem 1) and was presented as a block (stratagem 2). Only 8 sample questions are posed, but in practice, we would keep generating questions from this data until we ran out of ideas (stratagem 3).

The information provided to students as a "cover story" on which sample test items focus is shown in Figure 1. We should note that the text is one we use for college-level students already proficient in academic-style reading. Texts with similar content but simplified language can be adapted from newspapers and magazines and modified for use with students who have a more limited reading background. Students received five additional tables containing the raw data, numerical summaries, and the results of analyses (see Wild, Triggs & Pfannkuch (1994).

Background information

Monitoring for human exposure to environmental agents that might cause genetic damage and determining the type and extent of such damage is a subject that is commanding increasing attention. One method suggested as a possible measure of the genetic damage present in an individual's DNA is derived from counting the number of SCE's (sister chromatid exchanges) observed per cell. An SCE results from a reciprocal exchange of DNA between two spiral filaments that constitute a chromosome (sister chromatids). For each individual, a sample of cells is taken and the average of the resulting SCE measurements is recorded. This gives the value of the variable we call MSCE for that individual.



Dot plots for data on DISPERSION with superimposed LSD-display intervals.

One scientist looked at the average MSCE levels for four different racial groups and found no significant difference. However, other scientists have suggested that a measure of the variability of SCE measurements within an individual compared to that individual's average SCE level (giving a variable we call DISPERSION) may be a better measure of genetic damage for that individual.

Figure 1: Cover story for test items

Example 1: What to investigate next?

Suppose that you are listening to a conversation between five students about this problem. Which student would you agree with **least**?

(1) There appears to be a significant difference in DISPERSION levels between "Asians" and "Caucasians." I would like to check out whether this was due to a difference in the time of the cell scoring.
(2) We are not told about the genders of the subjects. I would like to check out whether the differences could be due to gender differences rather than racial differences.
(3) Are the smoking histories of the groups comparable? I would like to check whether differences in exposure to smoking explain the differences between racial groups.
(4) If you disregard the outlier in the "Black" group, only the "Asian" group stands out as different from the others. I would like to check out whether the "Asian" group are first-generation Americans as diet may be a possible cause of the difference.
(5) There being no real genetic differences between racial groups, I would like to check out whether these published differences result from the political beliefs of the researchers.

The habit of looking at a study and posing questions for further exploration is a crucial part of statistical thinking (Wild, 1994) and one that we really need to emphasise for future students. It is also an enormously difficult area to address with multiple choice. We appeal to our earlier adage, "Better to assess the important inadequately than the unimportant well." We do not expect this question to work well as assessment per se. Its main role is to say to future students using the exam in their studies, "Hey, this is important. Don't overlook it." It falls short in that we can only present ideas to be validated. We cannot ask students to come up with their own ideas. Our answer : (5).

Example 2: Nonsignificant versus no different

In this particular study and ten other similar studies, no significant difference was found between the sexes. Recently a researcher did find a significant sex difference in SCE levels and had her findings challenged because "everyone knew" a difference did not exist.
Which one of the following statements is **false**?

(1) The researcher who found a significant difference could not have used randomly ` sampled experimental subjects.

(2) On the basis of the previous studies, “everyone” had erroneously accepted the null hypothesis of no difference between the sexes.
(3) A real but small SCE sex effect would often not be detected by studies with small sample sizes.
(4) In the long run, the practice of “rejecting” hypotheses at the 5% level of significance results in 5% of studies investigating differences where none are present, coming to erroneous conclusions.
(5) Misunderstandings would have been less likely to arise if each of the studies had quoted confidence intervals along with the *P*-values.

We have taken an incident requiring statistical judgement and run items that may be causes or implications of the incident. This follows a pattern we use of confronting students with common misconceptions that arise in statistical thinking as a means of developing a critical attitude towards, and an understanding of, statistical aspects of published research papers. Most of this question attempts to put the student in the position of a research reviewer, to conjecture why and where possible errors of reasoning could have resulted in a “common belief” by scientists. We have asked for one false statement to be picked out from a batch of true statements (stratagem 7) with the true statements reinforcing important messages so that they act, at the same time, as a learning device. This question draws on more ideas than we would generally like in one question (stratagem 5), but this priority is overridden by the desire to avoid obviously stupid alternatives and stay close to factors at work in the incident. It works at the interface between three of the aspects listed under stratagem 4, namely: critiquing practical aspects of a study, interpreting data, and interpreting statistical ideas. (The multiple-choice format for communicating an answer to this question is imperfect in that no explanation of why the false statement is false is required.) Our answer : (1).

Example 3: How do we look for a relationship?

If we are interested in looking at the relationship between a person’s MSCE value and their DISPERSION value, we should look at:

(1) A scatterplot of MSCE versus DISPERSION.
(2) A stem-and-leaf plot of the differences between MSCE and DISPERSION.
(3) Side by side stem-and-leaf plots of MSCE and DISPERSION.
(4) A stem-and-leaf plot of the ratio, MSCE divided by DISPERSION.
(5) Side by side boxplots of MSCE and DISPERSION.

This question tests which plotting techniques give the desired information for a given situation. When data is produced it is important to know how to deal with that data and what techniques are appropriate for the formulated question (see stratagem 4). This question is much more focused than the previous one (per stratagem 5). Similar questions can be used to address the choice of modelling techniques. Our answer : (1).

Example 4: Reading and interpreting the plot

Which one of the following statements about the data on the variable DISPERSION (see Fig. 1) is **false**?

(1) Removing the observation with value 2.63 from the “Black” group would increase the evidence of a difference in mean DISPERSION between the “Black” group and the “Caucasian” group.
(2) The highest average level of DISPERSION appears to be in the “Asian” group.
(3) There is an outlier in the “Black” group.
(4) Removing the observation with value 2.63 from the “Black” group would strengthen any evidence of a difference in mean DISPERSION between the “Black” group and the “Asian” group.
(5) Ignoring the observation with value 2.63 in the “Black” group, the variability of the DISPERSION variable appears similar for the “Black,” “Caucasian,” and “Native American” groups, but looks somewhat larger in the “Asian” group.

In the list given under stratagem 4, this question corresponds to “interpreting data information.” Students have to know that the features being described here are best summarised in the plots of the raw data—no reference clues are given. Statistical jargon has been avoided to the best of our ability (stratagem 8). We are after an accurate reading of information in the plot and an intuitive understanding of how point patterns relate to formal evidence of differences between group means. We would often ask about features of plots without the link to formal inference being made here; that link was suggested by the data itself. An appreciation and intuitive understanding of variability is also reinforced. We have used a single false statement again, (per stratagem 7), partly because writing true statements about a plot like this is much easier than coming up with false but plausible statements, and partly because the four true statements used will help future students learn how to “read” data like this. An inadequacy of the question is that students are validating reactions to the plot rather than generating their own. Our answer : (1).

Example 5: Verbal understanding of a technique

Which one of the following statements about the application of one-way analysis of variance to the data on DISPERSION is **false**?

(1) The null hypothesis being tested is that all of the underlying mean DISPERSION values for all groups are identical.
(2) The *F*-statistic compares the variability between the sample means with the variability within samples.
(3) Large values of the *F*-statistic provide evidence of a difference between underlying true means.
(4) If the *P*-value is large, this implies that all the means are the same.
(5) A small *P*-value suggests that the data provides evidence of differences in underlying mean DISPERSION between some of the groups.

In the list given under stratagem 4, this question corresponds to “interpreting statistical ideas.” It tackles the big ideas of one-way analysis of variance and the F-test. It combines ideas about the hypothesis tested and the meaning of large and small *P*-values in the ANOVA setting.

Sometimes we would take two questions to do this. The *P*-value as a concept would have already been tested on its own in a simpler setting. This question is about what is special about one-way analysis of variance. The understanding of how a significance test is set up statistically, the interpretation of summary statistics, and what can be concluded from them are tested here. Jargon words are translated into terms that are as concrete as we could get them (stratagem 8). We also rely on stratagem 7 (all statements but one are true) to reinforce the main ideas about one-way analysis of variance. Our answer: (4).

Example 6: The purposes of formal tools

Which one of the following statements about analysing data is **false**?

(1) Plots can help tell us which methods of analysis are appropriate.
(2) It is good general practice to look at plots of data before performing formal analyses.
(3) We may use formal tests and intervals to help determine whether effects we have seen in our data are real or may just be due to sampling error.
(4) Even though we test null hypotheses of the form H_0: $\mu_1=\mu_2$, we would seldom, if ever, expect two population means to be identical.
(5) A *P*-value for testing a null hypothesis decided upon after looking at the data provides at least as much evidence against the hypothesis as it would have if the hypothesis had been decided upon before the study began.

In the list given under stratagem 4, this question corresponds to “what techniques should be used.” It is a general question about approaches to data analysis. It seeks to reinforce problem solving strategies and what can and cannot be reasonably concluded (stratagem 7). Our answer: (5).

Example 7: How is the P-value obtained?

If the *t*-test statistic for testing for a difference in mean MSCE level between “Native Americans” and “Caucasians” is 1.7, then the (2-sided) *P*-value for this significance test is:

(1) pr(2 *T* _ 1.7)   (2) pr(*T* _ 1.7) (3) pr(*T* _ 2 ¥ 1.7)
(4) pr(T _ 1.7)/2   (5) 2 ¥ pr(T _ 1.7)

This question examines the relationship between the test statistic and the 2-sided *P*-value in a non-traditional way. Other variants for testing the relationship between alternative hypothesis, test statistic, and *P*-value range from simply referring to a table to identifying shaded pictures. Our answer: (5).

Example 8: Interpreting a particular confidence interval

Suppose that we are calculating a 95% confidence interval for a difference in mean DISPERSION level between the “Blacks” and “Native Americans.” Which one of the following statements is **true**?

(1) If the interval is (-0.1, 0.9), then with 95% confidence, the underlying mean DISPERSION for “Blacks” is somewhere between being 0.1 units bigger and 0.9 units smaller than the mean DISPERSION for “Native Americans.”

(2) If the interval is (-0.1, 0.9), there is a significant difference in the means at the 5% level.

(3) If the interval is (0.1, 0.9), then with 95% confidence the underlying mean DISPERSION for “Blacks” is bigger than the mean DISPERSION for “Native Americans” by between 0.1 and 0.9.

(4) If the interval is (-0.9, -0.1), then there is no significant difference between the means at the 5% level.

(5) If the interval is (-0.1, 0.9), then with 95% confidence the underlying mean DISPERSION for “Blacks” may be smaller than the mean DISPERSION for “Native Americans” by more than 0.1 units.

In the list given under stratagem 4, this question acts at the interface between “interpreting data” and “interpreting statistical ideas.” The statistician has “analysed” the data and come up with a confidence interval, but what do these numbers mean? This question examines interpretation of the numerical values of the confidence limits for a difference between two means (in contrast to another question which would target what confidence intervals mean as a general concept). Communicating statistics by being able to translate statistical ideas into plain English is a goal of the course. Three of the statements manage to obey stratagem 8 (nontechnical language), except for the jargon phrase, “with 95% confidence.” Two statements make the link between confidence intervals and significance testing. We can thus expect some mixed messages in the item analysis for this question. Our answer: (3).

WHAT HAVE WE LEARNED?

Costs

The biggest advantage of forced-choice testing methods is that when a single test is given to a large number of students, faculty time is saved because the grading costs are negligible compared with other methods of assessment. A 65-question multiple choice paper takes us about 6 person-weeks to produce (including a rigorous checking process). A significant amount of that time is dictated by our desire to use fresh, context-rich data sets, which we would want to do irrespective of the form of examination. If our stratagems are followed, once the data file has been assembled most of the questions almost write themselves. We would put the additional cost of setting the examination in multiple-choice, as opposed to free-response and producing marking guides, etc., to be about 3 person-weeks and certainly no more than four. However, because of our large numbers of students, this 3-week fixed cost is overwhelmed by huge time savings on marking. It used to take us about 25 minutes to mark a traditional 3-hour examination script at this level (at that time we examined only mechanical tasks — verbal responses would take considerably longer to mark). Using the 25 minute figure, we estimate the marking costs for a traditional examination for our system to be at least 27 person-weeks, 9 times the cost of setting the exam in multiple-choice format. The additional cost is equivalent to one person committing 50% of her or his entire year to marking! We emphasise that the comparison here is just between multiple choice and another relatively inexpensive method, namely, the traditional 3-hour free-response examination.

Educational advantages

Other recognised advantages (Popham, 1990; Gronlund, 1993) are that a good deal of subject matter can be covered in a reasonably short period of time, and that marking is objective and consistent. We have found that the decoupling of skills forced upon question writers by the multiple-choice format has had some unexpectedly useful side effects.

The targeting of single ideas. Multiple-choice tests can form an information-rich source that is excellent for highlighting important ideas for future students. Having taught an idea or set of ideas, one can point very directly to that idea being assessed in the past.

Practising those parts of a skill which are causing problems. The breaking down of tasks into sub-tasks to construct multiple-choice questions can enable students to see which particular skills or parts of a skill are causing them problems and provide a means by which a student can devote additional effort to those troublesome areas alone.

Encouraging reflection and discussion about teaching. An advantage of the multiple-choice method that took us completely by surprise was that the discipline imposed by the format on question writers can improve teaching. Having to decouple skills and think through a whole series of options to write alternative answers for multiple-choice questions forced us to confront many of our own underlying assumptions and analyse in much more detail what it was that we really wanted to teach. Multiple choice alternatives tend to lay these things out for the whole world to see. Trying to write alternatives that assess big statistical ideas has increased the pressure on us to come up with ways of expressing those ideas concisely in plain English. Everyone on the team discusses and criticises each question *and all alternative answers*. The whole assessment process is completely explicit. Differences between teachers in preconceptions and interpretations are discussed at length with positive benefits for future teaching and future assessment. Any other forced-choice method can have these benefits.

Confronting students with common misconceptions. This is useful for refining the understanding of concepts. Once its value is appreciated, it can be done in any test format, but the nature of multiple-choice question writing forces teachers to think this way.

*Complete consistency in marking.*We get better quality control from a multiple-choice method than we used to get with long-answer questions. With long-answer questions, the marking scheme for a question tended to be read only by the person who made it up and then used it. The myriad of small decisions still to be made when marking a free-response question reflect the opinions of a single person. The credit the teacher should give for the thinking a student has done after taking a wrong turn (such as a simple computing error) on a long-answer question poses major problems, both in terms of equity and consistency. The big variation between different markers grading the same long-answer questions is well known, and there is even appreciable within-marker variation. Yet, marker variability pales into total insignificance when compared with the additional variability introduced by the practice of giving completely different tests to the students in different sections (streams) of a course. The “fairness” issues here should not be taken lightly. Students often end up in a particular section, or class, through a sequence of random events related to rooms and timetables—it is not a matter of informed choice. Therefore, they should not be penalised for ending up in the “wrong,” or difficult, section.

Some practical benefits for the exam itself. We used to find that with long-answer questions that yield partial credit, a single idea could accumulate credit across different questions throughout a paper. With multiple choice, we give one unit of credit in one place for one idea. With long-answer questions, mistakes or gaps in knowledge in early parts of a question can badly damage answers to the rest of a question.

Considering some of the criticisms

The criticisms made of multiple-choice testing are all valid to a greater or a lesser extent. This section moderates some of the criticism and outlines some additional attempts to mitigate the deficiencies.

"High level thinking." The idea that multiple-choice methods cannot test "high-level thinking" and is confined to low-level skills such as recall and application is overly pessimistic (Pandey, 1990). There is a great deal more analysis, synthesis, and evaluation required by the examples we have presented here than in many free-response items written by people who do not have our commitment to exploiting context-rich real data. It could be done better with free responses, but not without cost, particularly for those students who were stymied early in the process and thus lost the ability to show what else they knew or could do.

Entrapment by subtleties and fractionating knowledge. The nature of writing alternative answers can encourage a puzzle-writer's mentality whereby one ends up trying to trip students up with subtleties rather than examining big ideas. Because everything has to be broken down into small pieces, it can be easy to fractionate knowledge at the expense of interrelationships between concepts. However, we believe that these are pitfalls that question writers need to be aware of, and try to avoid, rather than being inherent flaws of the format. The statements we have used to construct examples 2, 4, 5, and 6, in particular, draw heavily on relationships between concepts.

"Multiple choice encourages rote learning." This is by no means inevitable. Even our mechanical questions (not presented to conserve space) rely very little on rote-memory. Teachers have to make conscious decisions about what is worth committing to memory and what can be obtained from reference materials when it is required. They must then provide the support materials that permit students to work in this way. We expect recall of key ideas only.

Tests used as a study resource. Because multiple-choice tests are so rich in information, we have found that some students use past tests as their primary study resource in preparing for future tests. This underlines how vital it is that test questions be focused on the most important parts of the course and not on obscurities.

Deleterious effects of time constraints. Criticisms such as "you only have two or three minutes per question" are oversimplifications. We gain from students only having to master one story to answer a large number of questions. The answers to some questions come almost immediately, leaving much more time for items that require more reflection. As with any other form of time-restricted test, it is important not to pack too many questions into an exam paper. We would far rather have students finish early than have them under severe time pressure, for such measures more accurately their in-depth knowledge. To deal with even stronger criticisms, we discuss areas where multiple-choice testing is perceived as totally deficient in the next three items.

The ability to perform a sequence of tasks unprompted. Multiple choice is clearly deficient here. We have to break down tasks into a series of subtasks to be examined separately. At each stage only the next step of logic, or a process, needs to be supplied, and then it need only be recognised. For largely mechanical tasks, it is often not practically important for later life to remember the sequence of steps precisely. One can look up the steps if and when one needs to use them. If they are used often enough, recall will soon follow.

Mechanical procedures are also prime candidates for automation. Take for example significance testing: The essentially human inputs, knowing what to test, being able to interpret the results of the calculations in terms of the original problem, and awareness of the need to check assumptions (e.g., about the distribution being normal) can be targeted using multiple choice. The graphical questions in our examples test accuracy of interpretation. This does fall short of what we would prefer: "What does this display tell you about the data?"

Language and communication. When responding to multiple-choice questions, students are forced to conform to the examiner's language in the formulation of the answer as well as the formulation of a question. Three aspects to this issue should be discussed.

- Fairness: Multiple choice is no less "fair" than free response. It is true that despite having its language polished by a whole team of people who are intent on minimising the possibility of misunderstanding, many alternative answers will be misunderstood by some students, and a small proportion will be misunderstood by large numbers of students. *All* communication is imperfect. Free response leads to unpolished one-to-one communications and therefore, in all likelihood, a greater probability of communication failure. A marker (grader) may cope with this by assuming that, whenever an answer fails to communicate itself to her or him, it is the student's fault. This is scarcely "fairer."
- Learning to communicate: The big problem with putting words into someone's mouth, as forced-choice testing does, is that it does little to help students develop their own voices. The best we can do with multiple-choice questions is to try to ensure that students have the ideas right and are used to seeing these ideas in simple language. This is one of the flaws of multiple choice that we pursue in the next subsection.
- Lack of information on student thinking: Pfannkuch has been researching the reasoning processes students use to form their answers by using extended, open-ended interviews of volunteers. Her research has shown that students go through very rich reasoning processes as they consider alternative answers and that these processes can display a great deal of statistical thinking. With forced-choice testing, one only sees the final choice. The greater the amount of reasoning needed between question and answer, the more severe the problem. This situation is much easier to circumvent with mechanical questions than it is with interpretation questions (simply break the task down into smaller components). This, of course, in no way obviates the need for the interpretation questions. Our hope is that our research into the thought processes of students may suggest new items that get closer to the ways students think, ways which can help steer that thinking in productive directions.

The random element. A much more obvious disadvantage of forced-choice testing is the random element introduced by students guessing answers. This problem becomes less serious as the number of questions becomes large and as the number of distractors becomes larger (we use 5 rather than 3 or 4—it is too hard to write more plausible answers).

What multiple choice cannot do

There are crucial elements of the statistical process that are impossible to experience and assess with forced-choice testing; in particular, anything requiring open-ended thinking. This often shows up in problem formulation (or question generation), in developing the ability to communicate, and in developing the ability to work or reason through substantial problems without step-by-step prompting. Forced-choice testing ties one into a world where everything is either right or wrong. There is no room for coming up with competing strategies and different perspectives. (Essentially the best one can do is to ask whether this is a promising approach or is it not, and is this one of the advantages/disadvantages of a certain choice or is it not.) More complex forced-choice test items such as assertion/reason items can do very slightly better (Frith & MacIntosh, 1984).

IMPLICATIONS

The authors have worked together on a team that teaches an introductory statistics course where there is more emphasis on fostering statistical thinking than on producing people who can perform a limited range of calculations expertly, and in which there is no emphasis on inculcating the ability to construct a mathematical derivation. Where the availability of resources and suitable coaches permits, we believe in the apprenticeship model of statistical education—students working on real problems from beginning to end under supervision—see the discussion of Romberg, Zarinnia & Collis (1990) on instruments of assessment and Wild, Triggs & Pfannkuch (1994) on assignment work (including a strategy for mimicking the investigative cycle within the confines of the traditional assignment and on strategies for enhancing communication).

Sheer weight of student numbers and a poor staffing ratio compel us to retain the traditional methods of forced-choice testing and regular assignments marked by student markers. Our challenge is to push the limits of these traditional methods and come as close as is possible to achieving the goals pursued by authentic assessment techniques described elsewhere in the book.

The focus of this chapter on traditional assessment techniques was multiple-choice testing. The following strategies were presented, illustrated, and discussed in detail:

1. Begin by collecting a file of real, context-rich data sets and stories.
2. Present the background, data, and a fairly large array of numerical and graphical summaries derived from it as a complete package.
3. Ask as many questions from the same story as possible, letting the story/data/situation suggest the questions.
4. Incorporate questions: critiquing practical aspects of the research; interpreting data-information and making inferences from it; interpreting and understanding statistical ideas; specifying what techniques should be used in a given situation; and performing mechanical tasks (e.g., calculations).
5. Break down tasks into subtasks and examine each subtask (alternatively, a coherent set of ideas) separately.
6. Provide enough information to prevent students wasting time on unproductive rote learning.
7. For questions about interpreting statistical concepts, or for “what to use where?” questions, bury one false statement amongst a collection of true statements.
8. Avoid statistical jargon in interpretational questions.

Only stratagems 5 and 7 are peculiar to forced-choice testing and only 7 to multiple-choice. The rest apply to almost any form of formal test, whether forced-choice or free-response, and to any level of the school curriculum. The above strategies are simply devices for getting as close to the reality of a statistical investigation as we can in the artificial context of a time-limited test.

Multiple choice does not allow students to experience some crucial parts of a statistical investigation and must be supplemented with other assessment methods. When circumstances dictate the heavy use of multiple choice, one should not simply accept the popular perception that it is an obstacle to better learning, but instead welcome the challenge to find creative ways to assess the skills that really matter. The disciplines imposed by the multiple-choice format on those teachers who are serious about trying to write good questions can have very positive effects on the interaction between teachers and on their teaching. The necessity to break down skills and ideas into component parts forces a deeper consideration of what it is that one is trying to teach. Disagreements between teachers over the meanings, usefulness, and correctness of questions and alternative answers provide profitable learning experiences for all involved. Whatever we do, we will always see students who are “collectors” who only learn isolated facts, “technicians” who want to isolate and learn well-defined procedures, and “connectors” motivated by a need to understand basic principles (terminology of Frid, 1992). Even using inexpensive methods of assessment, we have to find ways to create more students who are “connectors” if we are to have any hope of leaving them with something of value that endures for more than a few days after the final examination.