From **Gal, I. & Garfield, J. B. (editors). The Assessment Challenge in Statistics Education**. IOS Press, 1997 (on behalf of the ISI). Pages 153-164. ISBN 90 5199 333 1. The References section is a separate chapter and is available on the Internet at http://www.stat.auckland.ac.nz/~iase/publications/assessbkref.

12. Assessing Project Work by External Examiners

Peter Holmes

Purpose

Over a number of years the external examiners of a regional statistics course for 18-year-old students in schools and colleges in the United Kingdom became aware that the method of assessment was distorting the teaching and learning process, that the things being assessed were not the things that the examiners thought most important for the students to know. This chapter shows how the assessment methods were changed by adding in a compulsory project and reflects on the impact of this change on the teaching and learning of the students.

BACKGROUND

In England and Wales many 16 to 19-year-old students concentrate their studies on about three subjects, and at the end of two years of concentrated studies are assessed for what are called their Advanced Level General Certificate of Education (ALGCE). Some students will choose to specialise in mathematics, which may include some statistics, and a small number choose to specialise in statistics. Typically the syllabuses for these subjects are set by independent examining boards. The form of assessment and its implementation is also decided by these examination boards. Examination papers are set and marked by external examiners appointed by the boards. The only contact the students have with these examiners is through sending completed examination papers and other course work to them for assessment. Historically the main means of assessment, particularly in mathematical subjects, has been through three hour examination papers which the students have to answer on their own, without recourse to books or to each other or their teachers. In mathematics they may take one such paper in statistics. In statistics they will usually take two such papers. The results the students obtain in these examinations are highly influential in deciding whether, and to which, university they proceed. There is therefore a lot of pressure on them to get high marks in these examinations and other assessments made by these external boards.

From the external examiner's point of view it is clear that this scenario leads to a sort of game in which the student is under pressure to gain as many marks as possible, and this can be detrimental to encouraging a student to gain a deep understanding of the subject. Over the years examiners were becoming aware of the effect of the form and content of assessment on teaching and learning.

The challenge was to reconsider the impact of the assessment methods on students and their teachers. How could we change the forms of assessment to encourage students to gain a deep understanding of statistics, even if the student's aim was only to get as many marks as possible? (These two aims do not necessarily coincide, and in our case they certainly did not.) How could we change the form of assessment to encourage the right sort of teaching and learning?

For reasons given below, it was decided to make the students carry out a compulsory project as part of their course and to have this assessed. Also, for reasons given below, it was decided that the students should be able to choose the topic of their project. Since the ALGCE had a reputation of being an objective assessment of students' abilities any form of assessment had to be monitored and standardised, at least in the final analysis, by external independent examiners. The teachers or tutors of the students were not to be the final arbitrators of standard since it was considered that this was too subjective and arbitrary.

Since this was a new departure for students of statistics at this level, a number of questions were raised and had to be answered.

- Having decided on free choice projects, how could the external examiners assess the different projects?
- If any internal assessment were to be carried out by the teachers or tutors of the students, how could it be taken into consideration?
- Since the number of students entering these examinations was too large for all the projects to be assessed by a single external examiner, how could the marks of these external examiners be standardised?
- How do you cope with the limitations of external assessment when you can't see or interview the student?
- Since this was a new departure, what was the effect on students' learning?
- What was being assessed that previously was not being assessed?
- What showed up in this assessment that we had not previously known or had not been able to assess?
- How successful was incorporating project work into the assessment in improving the statistical understanding of the student?

I shall attempt to answer these questions based on my several years' experience as chairman of examiners and chief examiner for one of the ALGCE syllabuses in statistics that went down this road of broadening the assessment.

EXTERNAL ASSESSMENT OF COURSES IN STATISTICS

Curricular goals and principles

The position before the change

There have been ALGCE syllabuses in probability and statistics for over 30 years in England and Wales. The goals and principles of assessment reflected their historical development and were not properly geared to the goals of a current statistical education. Part of the difficulty was that they had always been seen as part of a *mathematics* qualification. Earlier syllabuses had

sections in pure mathematics and in applied mathematics (interpreted as mechanics). Since other subjects, such as economics, biology, geography, psychology, and social sciences were becoming more quantitative, there was a strong argument for replacing the applied mathematics section with a section on probability and statistics. It was thought that this would make the syllabuses more useful for students whose main interest was in these newer quantitative subjects. So new syllabuses were drawn up—but since they were part of a mathematics course they were drawn up by mathematicians. They were drawn up so that they could be seen as mathematically respectable, and so they concentrated on mathematical statistics. The three-hour end of course exam was taken as the appropriate assessment method, since this was taken for granted as being an acceptable means of assessing mathematics.

Several side effects followed from these basic decisions:

1. Questions in the examination would often concentrate on mathematical principles rather than statistical insight. For example, the student might be asked:

> A probability density function has the form $f(x) = ax^2+bx+c$ with 0_x<1.
> It has mean 0.6 and variance 0.1. Find a, b and c.

This requires no knowledge of statistics and only elementary knowledge of probability. It is essentially an exercise in solving three equations in three unknowns.

2. Questions which did try to explore some practical implications of statistical conclusions (such as asking the student to explain the practical implications of a given result being statistically significant) were notoriously badly answered, and also tended to carry few marks in the assessment. In general the form of the three-hour examination meant that very little emphasis was placed on the practical implication of the final result; indeed, most questions were based on false data invented for the question, so no true practical conclusions could be drawn in any case.

3. Because they were drawn up by mathematicians, the syllabuses were less useful than they might have been to the very students it was hoped would take the subjects. They did not link the statistics with the use being made in the other quantitative subjects.

4. A most important drawback was that the syllabuses, and particularly the assessment, did not encourage the students or their teachers to consider the statistical expertise required to carry out a complete statistical investigation. They did not encourage practical statistics or a consideration of real data and real problems. They were not developing true statistical expertise.

Reflecting on the position before the change

It was in considering these effects that we were led to ask whether what we were doing was best—and decided that it wasn't. We needed to sharpen up our aims in these courses and try to gear the assessment to reinforce the aims. It was clear that we needed to consider again what it meant to be a statistician at this level; what are the appropriate levels of statistical expertise we are trying to develop?

What is statistics? It is not a subset of mathematics, nor is it a set of techniques. Any definition must be broad enough to include all that statisticians do—applied statisticians in their different fields as well as academic statisticians. Statistics is a practical subject devoted to obtaining and processing data with a view to making inferences which often extend beyond the data. The statistician is involved with helping people make decisions in the face of uncertainty.

Including projects

Our reflections on these lines led us to decide that a compulsory project as part of the assessment would be a good thing. Since we wanted the students to be able to link this project with their own interests or with one of their other academic subjects, we decided to make the topic of the project open to the student to decide. We felt that through projects we could encourage a more rounded approach to statistics. There were also important things we could assess through the project work that we could not assess through the traditional three-hour exam. By a project we meant a piece of work undertaken by the student that would take about 40 hours of work, plus any time needed for data collection, that would start with defining a problem, collecting the appropriate data, analysing the data and drawing appropriate inferences. All this was to be presented in a written project report of about 15 pages.

The positive effects of projects

The team working on this change in assessment decided that there were many positive gains in requiring students to carry out a practical statistical project.

- Projects *put statistics in a context*, and make them more relevant. If the data arise in considering a particular problem, they have meaning. We do not have the problems of invented data; the data are throwing light on the situation. They are asking to be interpreted.

- Projects *give more motivation*. This is particularly so if the topics are chosen by the students themselves. They then do have a built-in desire to get insight into their problem. Sometimes the conclusions can be of real practical importance. I remember reading one project submitted by a student who had a weekend job at a local pharmacy. He was going to use data from this pharmacy as part of his project. The week before he was due to collect the data the pharmacist said that he was no longer required. The student persuaded the pharmacist to let him work for two more weeks (unpaid) whilst he and his friend collected data. They collected data on the number of customers, waiting times, the number of people who looked into the store but did not come in because it was too full, the amount of money collected in payment by the student, and so on. They then reworked these data on the assumption that the student had not been there, making a few other assumptions (usually in their own favour). At the end the report showed that the pharmacist would be losing money if he did not employ the student. As a result of showing the pharmacist the report the student got his job back. The student found real motivation in doing that project, and also completed a very good report.

- Projects *give a greater feel for real data*; their accuracy or otherwise; variability; reliability of conclusions; measurability. It is in collecting the data that you realise the decisions that have to be made, the way that some people are not necessarily answering the questionnaires truthfully, the possibility of bias in some of the measurements, and so on. It is this sort of feel for data that is most important when it comes to analysis and interpretation of the data.

- Projects *emphasise the application of statistics and its usefulness.* This can be a side effect in a larger group of seeing many students doing projects in many different areas. It is a good counterbalance to the impression that they can get from textbooks and from traditional examination questions.

- Projects show that *statistics is not solely mathematics.*

Projects or practical work?

An alternative to including project work as part of the assessment was to encourage practical work. The difference between practicals and projects can be summarised as follows.

A practical introduces or reinforces some particular theory; a project links a number of topics. Also, a practical has clear cut and limited objectives; a topic is investigative and more open ended.

In a practical the teacher defines the problem, decides the model, and decides the techniques; in a project the student defines the problem, chooses the appropriate model or models and decides the techniques. A project also requires a more substantial report to be written, while practicals are more easy to control than projects but do not develop global skills.

So, although practicals have a very important role, they do not develop the skills we wanted to encourage in the same way as do projects. We therefore chose to incorporate projects into the assessment process rather than just try to encourage more practical work. In fact we do try to encourage practical work as a precursor to projects, but that is not part of the assessment process, so is not necessarily done.

Anderson and Loynes (1987) give other good reasons for doing project work as part of a statistics course. They are writing from a university undergraduate and postgraduate perspective, but their points are still valid. It is interesting to see that they talk of skills that would not necessarily be considered statistical, but are part of good learning. They write of general non-technical abilities, such as working with others and communicating; general, partly technical, abilities, such as appreciating the ethics of a statistician and determining the aims of an investigation; abilities depending on technical skills, such as recognising which techniques are valid; and abilities depending on technical judgement, such as translating a real problem into statistical form. It should be obvious that it would be very difficult to assess these skills by the traditional three-hour examination.

Examples of project titles

One of the main purposes of including projects was to get the students to own the data, so we allowed them to choose their own topics. This did pose problems for assessment, and these are

considered below, but it meant that a wide variety of topics was chosen that reflected the wide interests of the students.

As an example, here are some of the topics that students chose in the first year of the new statistics course:

- Standing in line for school dinners
- Pupil preference for different chalk and board colour combinations
- The effect of “rounding” on regression lines
- Effect of soil conditions on the distribution of wild plants
- Human reaction times and learning ability
- Quality control of blood testing
- Consumption of gas in the UK
- Connection between rainfall and land height
- Predicting football results
- A game of golf
- Number of assistants required at a local shop
- Petrol sales at self-service and attended garages
- Mistakes in newspapers
- The effect of an overhaul on a canning machine
- Delays in flight arrival times at Manchester International Airport
- Analysing crime statistics
- A dental survey of children’s teeth in three schools

In general the better projects were those that were more tightly focused and that drew on the real interests of the student. Projects which tried to draw on large data sets, or were too ambitious, were less successful.

Principles which should guide assessment

The experience of examiners in the pure mathematics and statistics courses was that teachers were training their students to answer the questions being set in the three-hour examinations. Much time was being spent on teaching for this purpose, and it was distorting what the course was meant to be about. In particular it was not encouraging teaching that would lead to good statistical insight. I do not intend these comments to be seen as derogatory about teachers. If we as assessors imply that some things are important by the way we design our assessment, then it is reasonable for a teacher to try to enable the students to get as high an assessment mark as possible. The whole exercise becomes a game to get the most marks out of an examiner. This is particularly the case when, as here, the marks can be crucial for the student’s future career.

Many people have written more recently about what are good principles for assessment. The two major ones we took on board at the time of changing the course under discussion were (a) that the assessment should encourage good teaching and learning and (b) that the assessment should be objective and consistent.

We wanted to frame our assessment procedures to encourage good teaching and learning. Our assessments should seek to ensure that the content of the courses is right and that they develop the insights and abilities that are important. This includes process and global skills as well as particular techniques and topics. If we think that co-operation with others is important we must

include that in the assessment, even at the expense of having difficulty in evaluating an individual student's contribution to a joint effort. In technical terms this means that we may have to trade off some reliability against having greater validity. If we think communication skills are important for a statistics course at this level, then we should include them in our assessment.

The second principle in external assessment of the sort we were doing is that it should be objective and consistent. A mark scheme should be applicable to all legitimate types of project. If different markers are used they should come up with essentially the same marks.

Although the nature of the final assessment of the project was summative in that a mark or grade was given for the record, we wanted to encourage the project to be used as part of the teaching. So we involved the teachers with the project work so that they could do formative evaluation to help students learn before making the final evaluative assessment. Since the teachers were closely involved with the students doing the work, they actually had deeper insight into the nature of the student's understanding than was possible for the external examiners. So it was made a principle that the teachers do their own assessment and send these assessments and additional comments to the external examiners for them to use in their assessments.

Pragmatic examples using target knowledge or curricular goals

What is an acceptable project topic?

After deciding that a major project should be part of the assessment, we had to decide what sort of a project we would accept and hence encourage the students to do. After discussions with statisticians and teachers as well as with other examiners, we decided that anything which was a practical investigation using data would be accepted. It did have to be practical and it did have to use data. We wanted to allow students to use either primary or secondary data, or both, as appropriate. We did want them to do more than a trivial analysis of these data. We wanted the project to use some of the techniques that were part of the general syllabus content. This meant that we had to write our assessment scheme in sufficiently general a way so that all such projects would be included, and so that it encouraged a good use of the statistical techniques.

This decision did rule out certain types of projects that in other circumstances may have been quite legitimate. For example, historical projects such as *The rise and use of the Least Squares Principle in statistics,* and basically theoretical projects such as *The development of the theory of Markov Chains in probability* would not meet the guidelines. In fact, in the first year, the Markov Chain example was submitted. In its own terms it was a good project, but the teacher responsible had given poor advice to the student. To prevent the student from being unduly penalised, we, the external examiners, kept a close eye on the other papers done by this candidate so that she would not be unduly penalised. As it turned out, her examination papers were excellent and, even retaining the low mark for her project, she gained the highest grade on the assessment as a whole.

How much weight should be given to the project?

Another decision we had to make was how much of the total assessment mark would be given to project work. Too little and students and their teachers would infer that it was not all that important; too much and there would be political pressure (in the general sense) because of the perceived lack of objectivity in marking projects and in the possibility of students receiving help with their projects. Eventually we settled for 20% to the project and 80% to the final examination

papers as being politically acceptable. In terms of the effect on teaching and learning I think it would be better to move towards a 30:70 split and require more effort to be put into the project.

Marking the projects

The assessment had to be sufficiently general to cover all acceptable projects (as defined above) and had to be sufficiently precise so that teachers could carry out their own initial assessment and so that the team of external assessors could use it consistently.

The scheme set out to assess four qualities:

A The ability to select an appropriate probability model and set of statistical techniques for the investigation of the topic, and to design a plan of procedure;
B The ability to carry out a practical investigation in accordance with a plan of procedure;
C The ability to present the results of practical investigation in a valid and informative manner; and
D The ability to draw conclusions from practical experience and evidence.

The total marks were out of 40, divided A(8), B(9), C(10), and D(13).

To allow for the different levels of technique used in the different projects there were several general principles, such as:

- Mark generously for a difficult project involving difficult techniques, difficulty in collecting data, etc.
- Make appropriate allowance for the amount of direct help and guidance given (a form had to be filled in by the internal assessor and sent to the external examiner describing this).
- Discount work which is irrelevant to the project.

More detailed general headings were given under “A” to “D” to help with the assessment and allocate part marks.

For example, in “A” you were told to look for a description of aims and objectives, a description of the plan and design, the nature of the plan and its appropriateness to the aims and objectives, the probability models and statistical techniques used and their appropriateness. Later this was expanded for the benefit of the teachers who were doing the initial assessment. This included such statements as credit should be given for suitability of models, number of different models, and difficulty of the models used appropriate to the data. The description of the plan should indicate that some thought was given to such things as the data needed, how to collect the data, and how to use the data.

In “D” you were told to give credit for carrying out the analysis clearly and accurately, drawing valid statistical and practical conclusions from the data, discussing other possibilities and limitations of the project, and summarising the main points succinctly.

There were similar headings for the other two sections. External assessment followed internal assessment using the same scheme. The teachers made their assessment and completed a form to say not only what was their assessment but also what help the student had received. They had to

complete a declaration that it was the student's own work (except where due credit had been given in any references) and any other relevant comments.

The effect of the change

It soon became clear that the students were coming to the end of their courses with very different understandings and expertise than previously. Apart from the specific things we were able to identify in the projects, which are described below, there was a real difference in their answers to questions on the three-hour examination. In particular, the questions that had a request for a practical interpretation at the end were now answered much more sensibly. Whereas in the past we may have had an answer such as, "The result being significant shows that the data are significant," we were much more likely now to get comments relating to the context of the question, and even whether the underlying assumptions (say of the normal distribution or of independence) were correct.

Including project work meant that we could assess qualities in the student which would be very difficult in the traditional three-hour written examination. Even an oral examination would not have been as effective, and cost ruled this out anyway. Project work enables us to assess the student's ability to carry out a complete investigation from start to finish and to communicate the results. It reveals the ability to plan, work co-operatively, use limited resources effectively, adapt a plan as circumstances demand and complete it on schedule, and so on.

Pupils' difficulties as indicated by project work

After the scheme had been running a few years, we realised that we had found out much that we would otherwise not have known. We saw difficulties that the students were having that we had not known before. There were difficulties with the design of the project, with the analysis of the data, and with the written reports.

The difficulties in *project design* were associated with learning how to focus in on a sufficiently tight project that could be carried out in the time and with the resources available. We would find projects that did not succeed because they started with imprecise aims. At one extreme this could be the sort of project that worked along the lines of "let me find out all the data that I can that is at all relevant to my particular interest and see what I can deduce from these data." Such projects were not sufficiently well focused.

Other projects fell short because they were too limited in scope, or had repetitious designs; for example, a series of t-tests all of which were essentially the same.

Several projects could easily have been improved if there had been an appreciation of elementary design of experiments. For example: avoid confounding, use matched pairs to lower variability, use randomisation to eliminate other possible effects. An elementary consideration of the effect of sample size on the significance of results was also sometimes missing. Sometimes the samples taken were so small that it was impossible to detect any effect.

There were some surprising difficulties associated with the *analysis of the data*. It was not the purpose of the project assessment to look for details in the carrying out of the analysis. It would have been an impossible task for the examiners to check all the calculations. Even so, we did look to see that the summary figures and calculations bore some resemblance to the data. We expected most errors to be picked up during the normal consultation process between the student

and the teacher whilst the project was being done. Nevertheless it was surprising to find some gross errors.

There were errors in diagrams. For example, a fitted regression line would pass above all the points on a scatter diagram. There were also errors in the arithmetic. For example, the mean of a set of data was greater than the maximum value.

There was often a cavalier approach to data. This sometimes came from not knowing how to deal with outliers. Sometimes, genuinely, outlying values were dropped from the analysis because there was some good reason for thinking they were in error rather than just examples of high variability. Occasionally this process was taken to extremes with students automatically not using the highest and lowest values.

A continual problem was with students using significance rather than estimation. This was partly a function of the syllabus which emphasised hypothesis testing, but estimation and confidence intervals were also included and often this would have been a more informative approach to the data.

There was an occasional confusion between regression and correlation when students were not sure whether it was the strength of association or a prediction that was required.

The main difficulty that students had with their *reports* was a failure to draw valid conclusions based on statistical analysis, and prejudice showed through. One example was a study done on driving abilities of male and female students which, if anything, showed that females were better drivers than males. The student, a male, drew the conclusion that although the data seemed to show that females were better than males, everyone knew males were better.

Some students found it difficult to get the right balance between theory, detailed calculation, and summary in the report. The report has to show that the student has understood what is being done, but does not have to include large amounts of theory or repeated detailed calculation.

In spite of the above negative comments, the large majority of reports showed a much deeper understanding and ability to carry out a full statistical investigation and communicate the results than had previously been the case.

One area of concern was communicating with teachers. Because many teachers were involved, spread over a large geographical area, it was not possible to do other than have a couple of central meetings. Apart from these, communication was by printed material. The projects and the final marks awarded were returned to the teacher. This did help the teachers, over the years, to bring their assessment standards into line and to get experience in which subjects were and which were not likely to be good projects.

Realistic limitations and difficulties

The major difficulty was in devising the assessment to reinforce the learning principles. This has largely been covered above. There were other limitations and difficulties.

Initially there were not large entries to the examination—in the low hundreds. This meant that every project report could be collected in and re-assessed by the team of external markers. This small team standardised their marking by cross marking several projects and comparing their marks. Given the diverse subject matter they became quite consistent at this. This is important because such consistency is necessary if the assessment is to be reliable. A high level of reliability was expected for this examination as a whole. We all felt that the price of a little less reliability in order to gain greater validity (we were assessing the things we wanted to assess) was well worth paying.

Teachers seemed to be good at getting a reasonable rank order of the projects submitted from their school or college. Initially they were understandably less good at getting the absolute standard right. This was helped in succeeding years by the return of the projects from previous years and also by publishing a more detailed description of how marks should be allocated.

When large entries were received there was a plan to standardise by sampling. There was never any plan to standardise by reference to the results of the three-hour examinations since this would give extra weight to the examination, which did not measure the same things as the project. We felt that if we standardised against the examinations the students would concentrate on getting good marks in the examination at the expense of doing a good project, and we would be back where we started with bad assessment encouraging bad learning.

There were some difficulties involving how properly to incorporate teacher comment. The external examiners had to rely completely on the written reports from the teachers, which were fairly brief and variable in quality. In fact, only minor changes of marking were made if we received comments indicating large amounts of help being given. Our experience was that teachers were so anxious about giving too much help that they erred in the other direction.

Initially the teachers found their dual role difficult; they were both teachers and assessors. We tried to encourage them to see themselves as consultants, which was different from teaching. Their role was in asking questions rather than in giving answers. In practice the teacher's role changed throughout the course. Initially they were helping the students decide on a topic. They were helping them identify an area of interest, to see what questions could legitimately be asked, what data might be collected, and they were giving some sort of idea of the sort of answer the students might expect to obtain. Sometimes this took them out of a field of knowledge where they had any expertise. They had to develop the facility of asking probing questions and learning about the context from the student or other colleagues.

During the project their role was still as a consultant, and as an encourager to the student. They could also point out how the techniques they were teaching might be used in different projects, and use some projects as examples in their teaching. Before the final assessment the better students saw their teachers. At this time a teacher might spot major blunders and ask a question such as, "Are you sure that is right?" One difficulty was with students who did not keep in touch with their teachers, worked on their own, and submitted their projects at the last minute. Here there was no opportunity for the project to be corrected before assessment.

There is no doubt from our experience that the nature and form of assessment on the way the subject is taught and learned, and on what is seen by the students, is important. Including the project as part of the assessment did have many of the effects that we wanted.

IMPLICATIONS

It is clear from the effect of the changes we made that assessment does affect learning, attitudes, priorities, and the view that students have of the subject. We can not afford to continue with inappropriate assessment no matter how long the tradition behind us. Positively, if we orient our assessment towards assessing what is important, then this does change our students' attitudes to what is important (it also has a similar effect on some of the teachers). Rather than carry on as we have always done we need to think carefully about the purposes of our course and gear our marks to this purpose rather than vice versa. We must try not to let the assessment be seen as the most important aspect of the course.

If you are carrying out a change such as we did—giving a greater emphasis on project work and putting more responsibility on the student to choose a topic and on the teacher to supervise—then be prepared for opposition. It will come from people who will argue that the assessment is losing reliability and is less objective and less under the examiners' control. The response to this is that the total assessment is becoming more valid, we are assessing what ought to be assessed rather than what is easily assessed, and to some extent there has to be a payoff of reduced reliability to get greater validity; however, every effort has to be made to make the new form of assessment as reliable as possible. There will also be opposition from those teachers who do not want change, for they have become used to what they are doing. These teachers need to be persuaded of the correctness of what is being done and every effort made to quell anxiety and fear and to equip them for their new role in the whole process. In our case, if we had not done this the teachers would have transferred their students to other courses of a more traditional type. We provided meetings at which teachers could hear about and discuss the changes, ideas for possible projects, the reasons behind our changes—the change of attitudes we hoped to encourage, examples of projects, marks and other information. Once the new scheme had been started other in-service providers gave courses to help teachers in different ways.

Don't underestimate the effort needed to make such changes, but neither underestimate the harm that can be done by wrongly focused assessment and the good that can be done by well focused assessment. It may take a lot of time and effort, but the gain is worth it.