UNIVERSITY OF GLASGOW

DEPARTMENT OF STATISTICS

MSc.

“GOOD PRACTICE FOR FORMATIVE ASSESSMENT AND FEEDBACK IN STATISTICS COURSES”

Karina Paterson November 2008

ACKNOWLEDGMENTS

I am incredibly grateful to Professor John McColl, Department of Statistics, for all his valuable advice and recommendations. I would also like to thank Ewan Crawford, the Higher Education Academy's Research Fellow, for having made the computer based assessment system possible.

Many thanks also goes to the staff, postgraduates and students in the department that gave their own time to participate in this project.

Finally I would like to thank my partner for all their encouragement and support throughout this year.

CONTENTS

CHAPTER 3
Feedback Questionnaires

CHAPTER 4
Constructing Quiz System

CHAPTER 5

CHAPTER 6
Discussion and Conclusions

LIST OF TABLES AND GRAPHS

Page

CHAPTER 3

CHAPTER 4

CHAPTER 5

Abstract

Feedback to learners about their work is an important part of the teaching and learning process for any subject. Feedback should ensure students are clear on where they went wrong and what they can improve in the future. Without useful feedback students continue to make similar mistakes. However mathematical subjects such as Statistics appear to place less emphasis on feedback compared to other subjects. Statistics has made steady pedagogic progress and now uses a variety of assessment methods, but producing effective feedback for these methods has not made the same progress. This thesis investigates the feedback currently given to some Statistics classes for students who are studying Statistics as part of a degree in another subject, proposes a set of useable guidelines for producing effective feedback and reports on the creation and piloting of a multiple choice, computer-aided assessment system that provides immediate feedback to learners in Statistics courses.

The first chapter of the thesis discusses the background of the subject. Key features include the Quality Assurance Agency’s code of practice, which institutions should be following with regards to assessment, the National Student Survey, in which results for assessment and feedback are generally not favourable, and the various models for Statistics assessment suggested by Gal and Garfield in their book The Assessment Challenge in Statistics Education. An interesting thing about this book is that, though the whole book focuses on assessment, there is little mention of how to give feedback for any of the models.

Chapter two reviews the literature on feedback. This reveals that feedback can improve or impair performance depending on various factors. A summary is given of the most repeated guiding principles for constructing feedback. How students use feedback, including guidelines for receiving feedback, is also discussed. The final part of the chapter looks at the advice given for constructing

multiple choice tests and the lack of guidance for feedback relating to multiple choice questions.

Chapter three describes student questionnaires that were implemented in Statistics courses at the University of Glasgow to survey student attitudes to the feedback they received. A questionnaire and follow up questionnaire based on the guiding principles was piloted with a small group of second year Statistics students. Before issuing the follow up questionnaire, the way feedback was produced was changed in line with the guiding principles. When the questionnaires were compared, students were more satisfied with how quickly feedback was returned, the amount of feedback, the detail and the overall usefulness of the feedback after the intervention. The questionnaire was then adjusted to fit with a larger first year class. This included adding the Rosenberg self esteem scale to measure students self esteem. These results showed that the detail of the feedback given needs to be improved more than the amount. The most common reasons given for why the feedback was not detailed enough were that there was no suggestion for improvement, it was unclear where the mark was lost and the feedback was too vague. There may also be a relationship between students self esteem and the attention they pay to feedback. It appears those with a lower self esteem pay less attention to feedback. At the end of chapter three a briefing document is presented that can be used to help train markers. This is a summary of the guiding principles and includes good and bad examples of feedback.

Chapter four discusses the construction of a multiple choice testing system and the creation of specific tests for use in a level one Statistics course. First the chapter describes the piloting of another computer-aided assessment system called Model Choice. The results of this were very positive, with all students agreeing the system was easy to use and appreciating the immediate feedback. Next a similar system was created for use with the Statistics class for Psychologists and Social Scientists. Multiple choice questions were constructed for four of this course’s labs, on sampling and interval estimation, multiple regression, experimental design and categorical data. For each question, three incorrect options and a correct option were produced. Feedback was also written for each option explaining why the chosen answer was either correct or incorrect. Students getting the answer wrong first time were then given a second attempt. The literature on constructing multiple choice assessments was consulted during this process.

Chapter five focuses on piloting the computer-aided assessment system. The system was initially trialled with postgraduates and staff. The program received an excellent response and a group discussion revealed plenty of constructive ideas to improve the system. The program was then trialled with new third year Statistics students.

The final chapter summarises and discusses the results obtained to date and makes suggestions for further work.

CHAPTER 1
Background

1.1 Introduction

Statistics is a subject that is taught both to students who intend to specialise in the mathematical sciences and to students whose primary interest is in a wide variety of other disciplines. For example, when choosing Psychology as a degree, many students are unaware of the significant amount of Statistics work that is essential to the course. Quite a large number of the students appear to immediately dislike Statistics and think they will struggle with it. This is most likely because Psychology and Statistics seem to be very different subjects that involve diverse skills; students intending to complete a degree in Psychology do not expect to need arithmetical or mathematical skills but do expect to require advanced skills in writing reports and essays. Nethertheless students discover that they need to be able to master both of these sets of skills to be successful in Psychology.

One way to help students with this would be to improve the Statistics courses that non-specialist students have to take. Providing effective feedback on student work is an aspect that seems to require significant attention. There is evidence that feedback is an area that needs improvement in Statistics courses in general, not just those for non-specialists. (This is discussed in more detail later, with particular reference to the results of the National Student Survey.)

Feedback is an essential part of any learning process. If students are given feedback in an unhelpful manner they will be unable to improve or even to maintain the same standard of work. However the importance of feedback is sometimes overlooked in mathematics related subjects. This may be because the assessment in these subjects has historically been narrowly focused on tasks requiring technical skills and deductive logic, and the corresponding feedback will be different from that given on essays and similar work.

On the other hand, Statistics has progressed pedagogically and now uses a variety of assessment methods. Historically, many Statistics courses were only assessed using tasks involving deductive reasoning, similar to Mathematics. The story now is very different and assessment methods are used where an answer is not simply correct or incorrect (e.g. projects). Regardless of this, many teachers still mark work as if this is the case. It seems obvious that if there are guidelines for producing feedback for essay based subjects there should be some that are relevant for these more diverse Statistics tasks.

Fundamental to this thesis is producing a useable set of guidelines for constructing useful feedback. We will start by looking at what the Quality Assurance Agency's code of practice requires regarding assessment and feedback for students in U.K. higher education. We will also look at students' opinions of assessment and feedback throughout the country, as captured through the National Student Survey, to find out what standard is being set. We will show how this relates to the ways in which Statistics departments assess their students. Gal and Garfield's book "The Assessment Challenge in Statistics Education" (1997) has been very influential and gives a range of very useful types of assessment but, as we will show, it says very little about feedback to students.

1.2 Code of Practice: Assessment of Students

A code of practice was created by the Quality Assurance Agency for higher education (QAA, 2006) to assure academic quality and standards in higher education. Assessment of students is the sixth section and it aims to ensure

good assessment practice in organizations offering higher education. The general principles can be paraphrased as follows:
As bodies responsible for the academic standards of awards made in their name, institutions have effective procedures for:

1. designing, approving, monitoring and reviewing the assessment strategies for programmes and awards
2. implementing rigorous assessment policies and practices that ensure the standard for each award and award element is set and maintained at the appropriate level, and that student performance is properly judged against this
3. evaluating how academic standards are maintained through assessment practice that also encourages effective learning. (QAA, 2006)

Institutions should also be providing assessment that endorses successful learning. This can be achieved by planning a feedback loop into assessment tasks, setting extended assignments that involve researching, and peer assessed activities where students give feedback on each other's work. Furthermore the code of practice states that the amount and timing of assessment should allow for proper measurement of students' achievement of learning outcomes. Institutions will need to consider the other subjects students may take and give enough time for feedback to be put to use. With regards to marking and grading there should be clear assessment criteria which all students are made aware of. Students also need to be aware of the consequences that their achieved grade will have. This includes what they need to progress to the next module and how it will affect their end qualification.

The subsection on feedback makes it clear that this is very important when it comes to effective assessment. It is the institution's responsibility to provide appropriate and timely feedback to students.

“Institutions provide appropriate and timely feedback to students on assessed work in a way that promotes learning and facilitates improvement but does not increase the burden of assessment.” (QAA, 2006, p20)

Feedback should be present for all assessed work and should be given as soon as possible. It is emphasized that students require the feedback when they are most likely to pay attention to it. This means during a module and at its conclusion, but definitely not long after it is finished. To improve this, staff need to use their time efficiently and not give constructing feedback a low priority. Time pressures can be reduced by providing students with a collection of the most common comments or examples of work that were of an exemplary standard. The code of practice also suggests that students receive feedback from a variety of sources, including oral feedback, and that self assessment should be promoted. Furthermore students should be made aware of the types of feedback they will receive at the beginning of the course. This may also encourage teachers to keep to the high standard of feedback expected. It is also useful if the feedback refers back to the learning outcomes, giving the students a clear understanding of what is expected of them. Finally students should be praised for good work as well as given constructive comments.

The next subsection is staff training. The code of practice maintains that training should cover the different roles of formative assessment (where any mark for the task does not count towards the student’s final grade for the course) and summative assessment (where the mark is used to determine the student’s final grade), designing assessments, matching assessment tasks fittingly to the subject and awareness of cultural differences. It is also insisted that the vocabulary used in teaching and assessment should be the same. If this is not possible it must be made certain that the academic standard is not put at risk. In addition to this the code asserts that institutions evaluate and alter their assessment regulations regularly to account for changes in programme structure or external environment. Students need also to be made aware of their responsibilities when completing assessment. This includes them being aware of academic misconduct such as plagiarism and its consequences.

1.3 The National Student Survey

Whereas the QAA code of practice sets out standards for what assessment and feedback should be like, the National Student Survey (NSS, 2006) provides students with an opportunity to express their opinions on the assessment and feedback they have received, as well as other aspects of their higher education experience. Students give satisfaction scores in each of six sub-scales: teaching, assessment and feedback, academic support, organization and management, learning resources and personal development. They also give an overall satisfaction score. All these scores are in the range 1 – 5.

The NSS is now in its fourth year and the outcome for student's satisfaction with assessment and feedback has been relatively poor each time. Here are the statements regarding assessment and feedback to which students must respond.

- The criteria used in marking have been clear in advance.
- Assessment arrangements and marking have been fair.
- Feedback on my work has been prompt.
- I have received detailed comments on my work.
- Feedback on my work has helped me clarify things I did not understand.

The last two items seem to be the most problematic and generally receive the lowest scores.

In 2006, out of the six sub scales, assessment and feedback had the lowest overall mean score; 3.54 compared to the highest mean, 4.01 for teaching. Out of the 129 institutions in the survey 87 (67.4%) found assessment and feedback to be their lowest scoring sub scale. Looking at individual subjects, 66.8% of these subjects had assessment and feedback as the lowest scoring. For the subject group Mathematical Sciences, 61.3% of institutions teaching this subject found assessment and feedback to be the sub scale students were most unhappy with. Mathematical Sciences was not the poorest rated out of all the subjects, but this figure of 61.3% shows that it is a serious concern for a lot of students.

The overall picture of assessment and feedback in higher education is not a good one and institutions do not appear to be providing the academic quality that is expected by the QAA code of practice.

1.4 Assessing Statistics

In Statistics, the practice of assessment has been greatly influenced in recent years by educators such as Gal and Garfield, whose book “The Assessment Challenge in Statistics Education” (1997) illustrates a number of innovative models for classroom assessment. Whilst this book and related literature has encouraged higher education staff to use novel methods of assessment that examine a wide range of Statistical competencies, it does not offer them explicit or detailed guidance on how to give appropriate feedback on student performance.

The first model Gal and Garfield suggest is to use examples from the media in assessments. The media is full of topics that require statistical thinking and using these in assessment shows students how wide the need for statistical knowledge is. Two examples of how newspaper articles can be used for assessment are given. One example is to use a graphical representation from an article and ask questions on students’ understanding of it. Another is to give students a newspaper extract that includes information on a sample and a population and test their knowledge of the relationship between the two. There are many more ways media articles can be used and examples are easy to find since the topics using statistics are increasing. Gal and Garfield write in great detail about how to construct these questions but say nothing about writing the corresponding feedback. Since these were new and original methods of assessment it is even more disappointing that there was no advice on feeding back to students on their performance.

The next model is using a small group setting to assess problem solving abilities. The small group setting has been shown to be a very successful method for problem solving. The interactions within the group can provide more interpretations, contributing to the solution of the problem, although it is important to assess the different contributions. To assess problem solving and give a clear understanding of a student's performance, each individual's work as well as the group's work as a whole should be examined. This means the student's level of comprehension can be compared between these two assessments. The results of this comparison will need to be fed back to students but there is no mention of how. Careful feedback is necessary so that students know whether to concentrate on their individual or group skills. If students are only given feedback on one of these sets of skills the small group setting loses some of its power as an assessment tool.

Advice is also given on how to assess student projects. Assigning projects is a very useful method to test what students have learned in a practical context. Projects can be assessed during their completion in stages or once the project is finished. It may be more motivating to assess them in stages since it will give students an idea of how they are progressing. Two examples are given of staged assessment models. With both models, feedback can be given at the end of each stage. Verbal feedback will also be a useful contribution as it means students can respond right away and can reduce the potential for misinterpretation. This is as much detail as Gal and Garfield give concerning effective feedback. There is no information on what the feedback should contain or how the written and verbal feedback should differ.

A similar method of assessment to the project is a portfolio. This will present a student's achievements over time and may include both group work and projects. This is such an advantageous assessment technique because it displays multiple indicators of performance. For this technique to be productive, students must be aware it is being assessed, the criteria of what it should contain and the evaluation criteria. For this model Gal and Garfield do comment explicitly on feedback. The feedback given here must be practical for improving students' work and should give examples of how their work is different from what was expected. They even suggest anonymously using previous good and bad examples of portfolios, making it clear to the students what is expected of them. As this is one of only a few scattered references to feedback, it is unclear whether the authors are implicitly stating that feedback should be the same for all of their assessment models.

An assessment model becoming more common is based on the use of technology. Some of the most common are using computer software, computer simulations, multi media technologies and the internet. However there are some difficulties in using these for assessment. These range from a lack of appropriate resources to a shortage of training for teachers. If these problems can be managed then there are lots of advantages from using technology in assessment. It offers interactive learning benefits and using simulations means many concepts can be conveyed more clearly. One quite advanced method is dynamic external notations. Here the computer keeps a record of a student's activities while he or she completes a piece of work. In assessing these the teacher would be interested in the types of graphic manipulations they used, the files that were accessed to support their work, creativity used and the difficulties they came across. This type of assessment is quite different from the other types and surely this should be reflected in the feedback. However no suggestions are given on how to create feedback for computer assessment.

The last model Gal and Garfield describe is ‘how to assess on a budget’. This model is ideal for introductory classes where there are a lot of students. The focus of the model is multiple choice testing which has great time saving potential. There are some weaknesses in using this method (which we shall discuss at greater length in Chapter two) but Gal and Garfield claim that most of these can be overcome by using the following stratagems. The first step is to collect a file of real stories or data sets. All of the information, including numerical and graphical summaries, should be presented together so that students have to be selective of what is important. As many questions as possible should be asked from the same story as this will prompt new ideas. These questions should address evaluating practical aspects, interpreting data, explaining and understanding statistical ideas, identifying what techniques should be used and carrying out calculations. Tasks should be divided so that, if a student gives an incorrect answer, it is clear where the problem in thinking took place. However for this to be clear the appropriate feedback will also have to be given – a point which Gal and Garfield again fail to discuss.

All of the models discussed have their own advantages and disadvantages, and different combinations of these types of assessment are being used by Statistics departments in different institutions. It is striking, though, that every model described is lacking in detail on how to give feedback relevant to the assessment technique. Different assessment types appear to test different skills, making it difficult to apply the same procedures for feedback to all. Feedback is an essential part of assessment, but it seems likely that it has not been given enough importance in Statistics education.

1.5 Scope of Research

1) To review the relevant literature on assessment and feedback, especially any literature that deals explicitly with the subject of Statistics.

2) To establish general guidelines for effective feedback on performance in assessed tasks.
3) To determine how these guidelines should apply to various methods of assessment used in Statistics.
4) To investigate the quality of feedback in first Statistics courses at higher education level and pilot ways of improving this, if necessary.
5) To investigate the use of computer-aided assessment to give immediate feedback to Statistics students on multiple choice questions.

1.6 Ethical Approval

Ethics approval for the research reported in this Thesis was granted by the Faculty of Information and Mathematical Sciences ethics committee.

CHAPTER 2
Review of Feedback Literature

2.1 The Effect Feedback has on Performance

Feedback has been defined as ‘the actions taken by an external agent to provide information regarding some aspects of one’s task performance’ (Kluger and Denisi, 1996). Ramprasad (1983) also defined it as ‘information about the gap between the learner’s performance level and the reference level, which is used by the student to narrow that gap’. One of the important differences between these definitions is that Ramprasad’s includes the assumption that feedback is used by students and that this improves their work. In reality feedback, if used at all, has a variable effect on performance.

2.1.1 Improving Performance

Ramprasad’s definition, unlike Kluger’s, incorporates the idea of a reference level that the student should be aiming for. Confirmation of the importance of this comes from the work of Bandura and Cervone (1983), who gave undergraduates a strenuous activity to perform. They found that when feedback was negative, people would increase effort if the goal was clear, the commitment was high and there was a belief in eventual success. Young (2000) also found evidence of an improvement with extra work when students understood their assessment criteria. This improvement was also due to students appreciating the purpose of feedback, which means its definition is an important concept.

For students to improve their learning, the purpose of feedback must be recognized as facilitating their learning and not just as a judgment on their work (Maclellan, 2001). Students who tend to see feedback as a judgment are those with low self esteem (Young, 2000). To enable every student to benefit from feedback, Young (2000) believes that teachers should know every student's individual requirements. This is also the view of the constructivist theory of learning. This theory states that feedback acts as scaffolding which enables students to achieve more than they can alone. For this to be effective though, the feedback should be delivered at each individual's own level. Though this would be an ideal way to produce feedback, practically it would be very difficult.

Orsmond et al (2005) conducted a study with third year biology students using semi structured interviews and found that feedback could be used successfully to promote the managing of new knowledge into an already present structure of learning. This shows students can apply feedback productively to a learning framework. In Orsmond et al's study, five of the biology students carried around or saved the feedback they received showing evidence of the development of a learning framework.

When feedback improves performance it is often due to an increase in task motivation. However this result may depend on feedback being continuous. Several studies have shown that eliminating this feedback can remove the positive effect it had on performance. One of these reports was Komaki et al's (1980) study into the effect of training and feedback on the safety practices used in vehicle maintenance. As well as being continuous, in order to improve performance feedback should be revisited by the student. Smith (2007) found that students in an undergraduate geoscience course who revisited feedback or reanswered questions in response to the feedback, had an overall higher performance. These students received a mean exam grade of 79% compared to 70% for students who did not revisit or reanswer.

Learning has also been shown to be enhanced through instant feedback. Buchanan (2000) conducted a study with undergraduate psychology students and found those who had a web based formative assessment with specific feedback on each question, showed a superior performance in their end of course assessment.

2.1.2 Impairing Performance

Unfortunately there is a substantial amount of evidence which shows that feedback does not consistently improve performance. This evidence, however, has been overlooked by the majority of practitioners; e.g. Pritchard et al (1988) state that, “the positive effect of feedback intervention on performance has become one of the most accepted principles in psychology.” This view is also found in one of the most influential reviews in the feedback literature. Ammons (1956) states that knowledge of one’s performance increases learning, motivation and the level obtained by learning. However there are substantial problems with the studies Ammons cites. Many of these studies did not include a control group but only compared different types of feedback. This seems to be because the researchers have just assumed that feedback has a positive effect on performance. Ammons fails to mention that some studies found feedback could improve or decrease learning depending on what the participants were learning (Pressy, 1950). Furthermore Ammon’s evidence that feedback increases motivation, is taken from the finding that people have a positive attitude towards receiving feedback. These are not necessarily the same thing. However Ammons does recognize that feedback can decrease motivation if one is doing poorly. Similarly it has been shown that negative feedback, especially when it is recurring, produces a classical learned helplessness response (Mikulincer, 1990). This is the gradual, often false, awareness that the relevant activity is a hopeless endeavour.

Several experiments that concluded feedback does not improve performance were on memory recall (Fritz et al, 2000). These results cannot be generalized to feedback on more involved tasks. However a significant finding from Fritz et al’s study is that evaluating one’s performance during the task distracts attention from new information. This concept of attention is central to Kluger and Denisi's Feedback Intervention Theory (FIT). Kluger and Denisi (1996) constructed their FIT to predict the effects of feedback on performance. A key theme here is that performance depends on where attention is directed. Different variables will affect this, including personality and the way feedback is given.

To improve performance, feedback should be given that focuses attention on the task rather than the self. Classroom grades are shown to increase attention to the self and consequently produce no improvement in performance. This can be contrasted with specific comments which are task focused and affect performance positively (Butler, 1987). This has also been shown in a study with computerized feedback (Earley, 1988). The study involved sixty male and female magazine subscription processors working with allocated goals who received feedback from a supervisor or a computer. The computerized feedback was trusted more and resulted in a better performance than identical feedback from a supervisor. Kluger and Denisi's theory again explains this through attention. With the supervisor, some attention would be taken away from the task to assess the supervisor's intentions. Attention is also directed away from the task and towards the self when performance is compared with that of others or when feedback is made public (Kluger and Denisi, 1996). Similarly this will debilitate performance.

2.2 Guiding Principles of Feedback

Since there is evidence that feedback can have such a variable effect, it is essential to know how to construct good feedback. There are many different aspects to this and the literature's agreement on these elements varies. There seems to be six important principles that are mentioned frequently in the literature and which are generally agreed to be useful features of feedback. These are: balancing positive feedback (praise) and negative feedback (criticism) correctly, giving the right amount of detail, an appropriate quantity, being objective, timely and future oriented.

2.2.1 Balancing Positive and Negative Feedback

In producing feedback it is difficult to provide the right balance between negative and positive feedback. Negative feedback is more often given than positive and this is as high as 94% of statements being negative in one study with students studying English (Dragga, 1986). However a more recent study which analysed written feedback over a complete course of English as a second language, showed that positive feedback was quite frequent and that 44% of comments could be labelled as praise (Hyland and Hyland, 2001). Achieving the correct balance between these is a difficult process and is complicated by many factors.

Negative feedback itself can have damaging consequences but it can also be a method for improving students' performances. An important negative effect of criticism is the reduction in confidence and motivation in students (Taylor and Hoedt, 1966). For example criticism can be seen as a confrontation which serves to challenge a student's confidence. Many teachers are aware of this and keep criticism without any suggestion for improvement to a minimum. In Hyland and Hyland's (2001) study 76% of negative feedback was in some way made less severe. They found many teachers used imprecise quantifiers such as 'some' and 'little' to mitigate criticism.

Another method used to reduce the force of a comment is to phrase it as a question. This shows doubt on the part of the reader, indicating it is negative feedback but also that the writer should take action. However this is not always a useful way of providing criticism. Hyland and Hyland (2001) give instances where this type of subtle criticism goes unnoticed by the student. Therefore the student makes no adjustments and the feedback serves little purpose. On the other hand some researchers have found that the very nature of negative feedback leads some students to discredit it (Baron, 1993). This discrediting may also be due to the finding that negative feedback is perceived less accurately than positive feedback. Ilgen et al (1979) revealed this in their review into how feedback affects behaviour, concluding that negative messages may be distorted through a defence mechanism to guard one's self esteem.

Despite its shortcomings, negative feedback is a critical tool for helping students to realize and overcome their weaknesses. The purpose of feedback is to provide an accurate account of how good the work is. This would be virtually impossible without any negative feedback. Negative feedback is the only way students can find out what to improve. In a study at the Nottingham Trent University (Weaver, 2006) an overwhelming majority felt 'constructive criticism is needed to know how to improve.' This was 100% of the Art and Design students and 92% of the Business students. The crucial thing is that the criticism is constructive and delicately managed. It has been found that students are most motivated when their goals are not too difficult to achieve (Freeman and Lewis, 1998). This suggests all constructive criticism should be seen as attainable by the student. Nethertheless some students have shown no benefit from receiving criticism. Both Taylor and Hoedt (1996) and Gee (1972) showed no significant differences in their students' quality of writing after being given either negative or positive feedback on this area.

With regards to positive feedback there is evidence both for and against its effectiveness. Intuitively, praise for good work will result in an increased positive attitude (Gee, 1972). Many students can become pessimistic about their work and disregard the feedback if no positive comments are given. Giving feedback on what a student has done well is essential for knowing what to repeat in future work. Providing positive feedback also means students can plan ahead, as they know to give least attention to the areas they are already good at (Freeman and Lewis, 1998). These benefits mean students are really keen to receive positive feedback. Weaver’s study (2006) showed strong agreement from students that more praise should be provided.

However it has been found that continual exposure to positive feedback can prevent people from changing strategies when it is needed (Audia et al, 2000). Some students regard it as pointless feedback because it gives them nothing to build on. A significant amount of praise is felt to be dishonest and only there to soften criticism. Indeed Hyland and Hyland (2001) found that 20% of negative feedback was preceded by praise to produce balanced feedback. On the other hand, this can be a very useful way to provide feedback and it is often recommended to start with the positive. An ideal structure is the sandwich, where criticism is sandwiched between two pieces of positive feedback. It is also strongly recommended that the feedback avoids the word ‘but’ which can devalue the praise being given (Brockbank and McGill, 1998). If positive statements are seen as insincere it is doubtful they will be motivating. Brophy (1981) found that for positive feedback to be effective it needs to be informative and realistic. Nethertheless praise still tends to be much less specific than criticism. Teachers need to be aware of this and change their practice, since general statements are more likely to be seen as insincere.

2.2.2 Detail

Producing detailed feedback is another important guiding principle. General statements are of no use and all feedback should be specific (Brockbank and McGill, 1998). No benefit is taken from comments such as, “good piece of work” or “not good enough”. A study which interviewed students at Robbins University found widespread disappointment with how detailed the feedback was. Much of this focused on how little detail they were given on how they could improve their work (James, 1996). The situation appears even worse when the work is good.

General comments are more commonly found with praise and students find this very frustrating (Cowan, 2006). Both students and assessors want students to continue achieving high marks so it is difficult to understand why very little information is given on why assessments are good.

When feedback is made specific there is a reduction in students concerns over the fairness of their mark (Wilson, 1999). It seems that once provided with the correct information students can understand their mistakes and agree with the marker. Several researchers go as far as to say that specificity is correlated with learning (Vallacher and Wegner, 1987).

2.2.3 Amount

When considering how much detail to provide, the issue of what is a suitable amount of feedback to give will also arise. In the past it was quite widely accepted that the more feedback given the better (Ilgen, 1979). However more recent research suggests it is not quite as straightforward as this. Many believe that feedback should only focus on a few areas so that students know exactly what to change. This is especially true with negative feedback and for maximum benefit it may have to be limited to one or two areas (Brockbank and McGill, 1998). After this, many students switch off so it is better not to concentrate on insignificant or infrequent errors (Wilson, 1999). However Weaver (2006) discovered that 96% of Business students and 75% of Art and Design students felt they were not provided with enough feedback.

A very interesting finding (Vollmeyer and Rheinberg, 2005) with regards to how much feedback is useful, is that performance improves with no feedback being given and only the expectation of receiving it. This is thought to trigger a deeper processing of learning. This may be because students work more thoroughly if they know teachers are checking their learning outcomes. On the other hand, some believe the prospect of receiving feedback provokes fear in students so that they will not be able to respond positively to the remarks made to them (Brockbank and McGill, 1998).

2.2.4 Objective

Another important principle for constructing feedback is to make it objective. The best way to achieve this is to have both the handed in assignment and the feedback on it evaluated by someone other than the marker (Hirsch and Gabriel, 1995). This would not be necessary for every assignment; it would only need to happen randomly to ensure markers kept their feedback unbiased. Markers can also be objective by phrasing their comments as their own view. This shows their feedback may not be a universally agreed view and takes accountability for what they are saying (Brockbank and McGill, 1998). This is also a useful method for softening criticism. It acts to decrease the authorative gap between teacher and student. This may mean that students are less threatened and more willing to take the feedback on board (Hyland and Hyland, 2001).

2.2.5 Timely

The most repeated principle for effective feedback in the literature is that it must be timely. Feedback can never come too soon and should be given as soon as possible after the work has been completed. There is some evidence, however, that feedback given as work is being carried out increases anxiety and consequently harms performance (Wise et al, 1986). It may be that learning of one’s failure during performance impairs the rest of the individual’s work. Researchers have suggested this is because attention is aimed towards the self and not the task (Kluger and DeNisi, 1998). However this evidence is only relevant to feedback provided during performance and the student should always be receiving their feedback before their next assignment so it can be put to use (Hirsch and Gabriel, 1995). If the delay in receiving feedback means the student is on a different part of the course, it is unlikely any real attention will be paid to the feedback. Hartley and Chesworth in 2000 found 59% of students felt feedback was given too late to be helpful. Even if attention is paid, very few students will ask questions about their feedback if the delay has been long (James, 2000). The principle of timely feedback is thought to be even more important for first year students. For these students it is essential that, before more work is completed, they know what to change or are given the confidence they deserve (James, 1996).

2.2.6 Future Oriented

Finally feedback should be relevant to both the current assignment and future work (Hirsch and Gabriel, 1995). The principle of feedback as 'feedforward' is essential to learning oriented assessment (Carless et al, 2006). This means feedback should have clear implications for the current and future tasks. Therefore feedback should not be limited because the work is a final draft. Even though nothing can be changed on the current work the advice can be taken to forthcoming assignments. The marker should be doing more than justifying the grade for the current project. If feedback is written with this principle in mind recurring problems should be evident to the student.

2.2.7 Other Aspects

There are a few other principles that do not appear so commonly in the literature. Firstly some researchers suggest that feedback should not just give the answer, but should promote thinking in students (Hirsch and Gabriel, 1995). Some also believe the feedback should be both relevant to the assessment criteria and the individual student. To be relevant to the individual, this would include taking into consideration previous work. Another principle occasionally seen in the literature

(Carless et al, 2006) is that feedback should be a two way process and encourage dialogue between student and teacher. Finally it may also be important to focus on behaviour rather than the student and the focus should be on changeable behaviour.

2.3 How Students use Feedback

2.3.1 Forms of Usage

Orsmond et al conducted a study in 2005 to discover students' specific uses of feedback. The study involved third year biology students and consisted of semi-structured interviews. It identified four main areas of utilization and another two which were found to be occasional uses.

Firstly, many students use feedback to motivate them. This motivation can arise from both positive and negative feedback:

> "I believe that feedback should be critical and praiseworthy."
>
> (Student 4, Orsmond et al, 2005)

Positive feedback makes students feel more confident and can encourage them to keep working at the same high standard. Negative comments on the other hand will motivate students to improve their work. For this the criticism will need to be constructive and suggest improvements. However the reasons for becoming motivated did vary for some students. One student felt negative feedback motivated them to work better in order to prove people wrong. Another respondent became encouraged to approach lecturers for more help due to the feedback they received.

The next area where feedback is used is to enhance learning. Feedback is used as a guide by students to develop their work. This is the main goal of feedback and Orsmond et al (2005) found significant evidence of this from their students. Some students had experienced tangible improvement through their marks increasing. In addition it appears that feedback is often generalisable and can be used to improve learning in other subjects:

“Feedback helps with other modules and exams; you can avoid making the same mistakes.”

(Student 9, Orsmond et al, 2005)

The students also discussed how they would progress with their learning if there was no feedback. Most students found this inconceivable:

“Oh God it would be terrible. I would be working blind.”

(Student 11, Orsmond et al, 2005)

The majority felt there would be no way of improving without feedback.

Another common use of feedback is to enhance reflection. This means the students deliberate more on their work in connection with the feedback. They may approach their assignment in a different manner or combine and compare all their feedback.

The last main area is the use of feedback for clarification. Orsmond et al (2005) found that feedback gave a lot of students a clearer understanding of why they achieved their mark. Feedback also helped them comprehend the assessment criteria better.

“I think feedback helps you know what is expected at a given level.”

(Student 11, Orsmond et al, 2005)

Furthermore students gave suggestions on how feedback could clarify more. The main idea here was that there should be feedback on how students are doing overall as well as specific feedback for assignments. Students want to regularly know how their work is affecting their final grade.

There were two other uses of feedback which may impact the four already discussed. Several students seemed to apply feedback to enrich their learning environment. This was especially noticeable with students who received information before an assignment. This meant they could use this information to give context to the feedback they receive. Finally students obtained more use out of specific feedback. The majority of students preferred clear direction and suggestions in their feedback.

> "Feedback is easier to use if you have to do something mechanical, like being directed to specific resources."
>
> (Student 8, Orsmond et al, 2005)

As well as the areas revealed by Orsmond et al (2005) several sources maintain feedback is used to assess development and prepare for forthcoming learning (Cree, 2000). However the evidence for these claims is very limited. It is also unclear whether Orsmond et al's (2005) study really shows that students use feedback in these specific ways. The study only interviewed 16 students and it is unlikely this is representative of all students. The methodology for the study was a semi structured interview. However what a student says and the way they actually use feedback are not necessarily the same thing. Crisp (2007) conducted a study using undergraduate social work students and found little support that feedback leads to changes in subsequent work.

2.3.2 Engaging with Feedback

Some researchers claim the majority of students are primarily concerned with the grade they achieve and feedback is only of supplementary interest (Smith and Gorard, 2005). Evidence for this comes from a study by Higgins et al (2002). They found that 39% of students spend less than five minutes reading feedback and 81% spend less than fifteen minutes.

On the other hand, it has been suggested (Fritz et al, 2000) that the students who use feedback are those who can actively engage with it. One key aspect to this is that students need time to incorporate feedback before moving on. Fritz et al (2000) completed some interesting work relevant to this idea. They found that when trying to recall a passage subjects performed worse when their attention was focused on the mistakes they had made. The suggested explanation for this is that assessing your memory of the passage and taking note of the mistakes, made concentrating on new information more problematic.

Therefore as well as the content of the feedback being important, how the students interact with and understand it also matters (Orsmond et al, 2004). This is thought to depend on student's previous experiences (Ramsden, 1992) and their intellectual development (Perry, 1970). If students have a different understanding of the purpose of feedback to their teachers it will appear they are unable to utilize feedback.

Hounsell (1987) found that students adopt two main approaches when it comes to learning. These are deep, where the student concentrates on meaning, and surface, where the student orientates towards structure and content. This suggests it would be productive to introduce the roles of feedback through teaching. Using feedback productively is a skill and to develop this skill, students need to become aware of the wide range of uses of feedback. Interestingly very few students receive guidance on how to understand and use feedback. Weaver (2006) found, from studying Business and Art and Design students, that on average half of students have received no direction on using feedback. From the 50% who did receive guidance only 14% received this at university. The other 36% learned how to understand feedback before university or on their own initiative. These are worrying statistics since it has been shown that receiving feedback is not the same as being able to act on it.

2.3.3 Guidelines for Receiving Feedback

It appears that, as well as teachers being trained on how to produce feedback, students should learn how to receive it. The key ideas here can be seen as guiding principles similar to those for giving feedback.

The first skill students should possess is being able to clarify the feedback they receive. The student should be clear about what the feedback is saying and this may mean contacting the marker for a full understanding. With oral feedback it is useful for students to repeat their interpretation of the feedback to save any misunderstandings (Brockbank and McGill, 1998). Students should not rush into a response to the feedback. This could lead to the student being unnecessarily defensive and rejecting the feedback. They need to be able to accept constructive criticism as a useful tool and not as a personal judgment (Freeman and Lewis, 1998).

Another important guideline is students should be able to approach teachers for feedback. This may be because they did not receive enough feedback or that the feedback they received was unhelpful (Brockbank and McGill, 1998). It is also a good idea to have more than one perspective on a piece of work. This could come from another teacher or someone from the student’s peer group (Freeman and Lewis, 1998). This will be especially useful when a student disagrees with some of the original comments.

Finally and most importantly students should always be responding to the feedback. The majority of the time this will be taking direct action on their work but this can also mean approaching a teacher for clarification if what action to take is unclear (Fry et al, 2003). No feedback should be entirely ignored.

These guidelines show that both student and teacher have a responsibility when it comes to producing effective feedback.

None of these findings were the result of research into effective feedback for statistical work. When searching the literature, practically nothing could be found explicitly on how to produce feedback for Statistics. This is very difficult to believe since feedback is equally important for all subjects.

2.4 Computer Assessment

Glasgow University's Statistics Department is developing a form of computer assessment called Model Choice. This is a multiple choice test which gives students immediate feedback to their response. Therefore how to produce helpful feedback in computer assessment is an important issue here. However the literature on computer assessment does not discuss feedback in great detail. This is surprising since many practitioners are aware of its importance. One study into higher education observed that developing feedback for multiple choice and similar tests would be very beneficial (Gipps, 2005). They comment on how powerful a learning instrument it could be if feedback was automated while preserving quality.

One of the advantages of computer assessment is that it can deliver instant feedback. As already discussed above, this is more beneficial to students than when there is any delay (Dempsey et al, 1993). Students also have a very favourable opinion towards this type of feedback. Thelwall (2000) conducted a study into a randomly generated computer assessment. The students were from Wolverhampton University and were studying Statistics. Originally the test did not give feedback but after many students' requests, this was added to the assessment. The majority of students felt the feedback was useful (91%). The feedback can even be printed off and used for revision. Furthermore the computer assessment actually encouraged students to revise more. It seems to motivate students much more than other types of assessment. This is thought to be due to some students being motivated by a desire to accomplish high marks irrespective of the context rather than by a desire to learn the subject for its own sake, but it is still a benefit of this form of assessment.

Thelwall's is not the only study to find an increase in motivation from using computer assessment. In 2003 Blayney and Freeman found that 82% of students felt more encouraged to persist with their studies because of computer assessment. The 2005 Pass IT project that investigated the use of IT to support assessment in Scottish schools and colleges (http://www.pass-it.org.uk/resources/9500_pass-it_8pp_250705hr.pdf) established that this was due to the more interactive resources. Evidence for this also comes from a study using first year geography students at the University of Plymouth (Charman and Elmes, 1998). When this department changed the assessment for the module to become computer based, it was found that student's interest in the module increased. More students agreed with the statement 'I found this module stimulating and interesting.' This study's questionnaires also showed that 56% felt computer assessment was an improvement on other types of assessment and 85% that the assessment was fair. Another interesting finding is that students' marks appeared to increase with the introduction of the computer assessment. The average mark improved from 54.1 to 56.8. However this is only a small change and the minimum and maximum marks did not alter with the change of assessment. It seems that the computer assessment had the most benefit for students who were on the borderline of passing the module. With regards to feedback Charman and Elmes (1998) found that 68% of students indicated that it was adequate and relevant. This may be because the assessment gave specific feedback to each response and it is frequently found that the most helpful feedback gives precise and detailed comments (Black and William, 1998).

Crisp and Ward's (2007) study into producing a scenario based computer assessment also focused on producing useful feedback. The scenarios were based on classroom situations and the assessment was for trainee teachers. For each question, feedback was given that included the correct answer with reasons and, if necessary, why the answer was incorrect. Participants completed a questionnaire after the assessment to give their opinions. The results of the questionnaire showed that the average score for trainee primary teachers who felt the ‘feedback given was helpful’ was much greater than for the trainee secondary teachers. Further interviewing revealed that many participants felt there was not enough feedback. In response to this, more links were added for websites relevant to each scenario where students can learn more on their weak areas.

As well as providing feedback to students, an important strength of the computer assessment is how it can provide in depth feedback to the tutors on students’ performances. A disadvantage of this type of assessment involves the actual implementation of the test which requires the provision of a large number of secure browsers in an environment where the identity of candidates may be checked. Furthermore with the layout of the majority of computer workstations it would be fairly easy to view a neighbour’s screen.

2.5 Constructing Multiple Choice Tests

Multiple choice testing now has a long history of use in many subjects. In Statistics, its use was championed by Gal and Garfield (1997), who discuss its advantages and disadvantages in their book, ‘The Assessment Challenge in Statistics Education’. Multiple choice testing has many benefits with the biggest being saving staff time. The other advantages include that a large amount of material can be covered, marking is objective, single ideas can be targeted, students can identify exactly where they need practice and common misunderstandings can be highlighted.

It may also be that some of the common criticisms are unwarranted. One criticism is that multiple choice cannot test high level thinking. However examples can be constructed that require analysis, synthesis and evaluation and not just recall and application. It is also said that this method encourages rote learning. On the other hand support material can be provided which should prevent students from this type of learning. Another unjustifiable criticism is that time constraints are too harsh. It appears that, on average, only two or three minutes are allowed per question. However some questions will be answered almost immediately, leaving more time for thinking on other questions. Nethertheless there are some aspects of statistical assessment which multiple choice cannot capture. The most significant feature missing is the ability to test open ended thinking. There is no opportunity for students to express different perspectives.

Most authors, including Gal and Garfield, fail to see the need for feedback with multiple choice testing. However students will benefit more if they know exactly why their answer was incorrect. Part of this thesis describes a piece of work to introduce computer based multiple choice testing that gives immediate feedback to Statistics students (see Chapters four and five). The program was conceived as giving students two attempts at each question with feedback for both correct and incorrect answers. A program designed on these lines was expected to be effective as a learning tool as well as for measuring students' abilities.

Constructing multiple choice items is more complicated than it initially appears. How to produce a high quality assessment will depend on the purpose it is serving. There are two main purposes examination can have. The first is that testing students serves as a learning device and the second is that the test measures learning.

If the main aim is measurement then it is important that the test items satisfy two concepts; Difficulty and Discrimination. With regards to difficulty this can be measured using the Difficulty Index (DI). The DI is the percentage of students answering that question correctly (Baldwin, 1984). This is important because an assessment is not measuring any ability if every student is answering the question correctly. This is also true if all the students are answering incorrectly. It has been claimed that the optimal DI is 50% as this maximises the standard deviation of scores on the item, which can be seen as obtaining the maximum information from the question (Nunnally, 1970).

The other concept, item discrimination, is even more essential than the DI. This involves ensuring that the students answering correctly are also the students who understood the concept and vice versa. This is calculated by correlating the correct and incorrect answers with the total exam score (Baldwin, 1984). Ideally the correct answer should have a high positive correlation with the total score and the incorrect answers a high negative correlation with the total score. These two concepts are obviously crucial for effective test items but what is not so clear is how to construct items in line with these concepts.

Baldwin (1984) was one of the first to construct guidelines on how to write multiple choice test items. He gives eight practical pieces of advice and even though these are aimed at accounting education most of the recommendations can be generalized.

- Each question should assess an explicit concept. He suggests deciding on the principles you want to test and then constructing a specific item for each one. After the question is completed it should be clear that there are certain incorrect answers for commonly misunderstood concepts.
- Avoid questions that are too easy or too difficult. Research has shown no benefit of easy questions to build confidence (Howe and Baldwin, 1983).

- Avoid unnecessarily long questions or options. Short and straightforward questions measure the actual concept being tested more effectively.
- Use a mixture of different types of question i.e. conceptual and problem type questions.
- When testing a numerical concept use simple numbers. This means the question is testing the principle instead of calculator ability.
- Incorrect answers should be chosen carefully. Frequent use of ‘none of the above’ and ‘all of the above’ should be avoided. The incorrect answers should seem plausible to students who have not mastered the principle.
- Have an equal allocation of the correct answer over A, B, C, D and E.
- Avoid the possibility that the correct answer can be reached through incorrect logic. The example given for this guideline relates to accounting so it may not be easy to generalize.

(Adapted from Baldwin, 1984)

The majority of these guidelines seem intuitively like good practice but there needs to be evidence to support them. A number of researchers have conducted reviews of all the guidelines in the literature. In 2002 Haladyna et al looked at how valid 31 of these guidelines were. Two sources of evidence were used; the collective opinions of textbook authors and empirical research. The following guidelines are cited and supported by over 70% of texts and all the evidence for them is in agreement.

- Each item should display specific content.
- Items should be based on important principles in the course.
- Use novel material.
- Use clear directions.
- Ensure there is only one correct answer.
- The length of options should be roughly equal.

- Phrase options positively.
- Avoid clues to the correct answer.
- All incorrect answers should be plausible.
- Use common errors in incorrect answers.
- Use simple vocabulary.
- Avoid ‘all of the above’ as an option.

(Adapted from Haladyna et al, 2002)

The next set of guidelines were those that were supported unanimously by the authors who mentioned them, but were not mentioned at all in a significant amount of work.

- Keep items independent of each other.
- Avoid being too specific or too general.
- Avoid opinions.
- Edit and proof the items.
- Use correct grammar.
- Minimize reading for the student.
- Avoid window dressing.
- Vary the location of the correct answer.
- Keep options independent.

(Adapted from Haladyna et al, 2002)

For the remaining guidelines conflict existed either between authors or the empirical research.

- Avoid trick items. The authors collectively agreed with this principle but the only study into trick items revealed some interesting results. Roberts (1993) found when testing introductory Statistics students that they were unable to differentiate between trick and non trick items. Moreover even though the textbooks support this guideline it was uncited in 33% of work.

However Haladyna et al's conclusion is that this guideline should be supported.

- Format the item vertically not horizontally. With this principle authors are split between the two presentation methods. There is also no evidence that one format is advantageous over the other. However the review continues to defend the vertical presentation.
- The central idea should be in the stem (i.e. the initial statement of the question, before the optional answers are presented). Downing et al's (1991) research into focused stems discovered no difficulties with unfocused stems with the essential idea in the options. Nethertheless this guideline is still good practice.
- Phrase the stem positively. When analyzing citations 23% of these had no issues with phrasing stems positively or negatively. Furthermore 75% of the research into this guideline showed no difference in difficulty that was caused by wording the stem negatively. Haladyna et al conclude negative phrasing should be used with caution and when used the negative word should be highlighted.
- Have as many plausible incorrect answers as you can. 70% of textbooks agree with this guideline, 26% do not give any indication of how many distracters there should be and 4% think there should be a limit to the options. The issue of how many options to create is the most researched out of all the guidelines. These studies have found mixed results with some revealing that reducing the options reduces the difficulty (Rogers & Harley, 1999) and some that it increases the difficulty (Cizek and Rachor, 1995). In 1993 Haladyna and Downing assessed four multiple choice tests and found that only 1-8% had three successful incorrect answers with 67% having one or two plausible distracters. It was concluded that two incorrect answers is adequate for the majority of items since it can be difficult to produce three equally plausible incorrect answers.
- Options should be in a logical order. All textbooks citing this guideline supported it. However a study into Mathematics tests revealed no

difference in difficulty depending on ordering (Huntley & Welch, 1993). Regardless of this evidence the majority support a logical order.

- Options should be homogeneous in content and structure. The research again found no definite evidence for this guideline (Downing et al, 1991). However there is a great degree of agreement among authors so the principle continues to be supported.
- 'None of the above' should be used with care. This guideline produced the largest divide between authors. 44% agree with the guideline, 48% think it should never be used and 7% of textbooks do not cite it. All of the studies into this option stated that it increased difficulty. It is advised that new writers completely avoid using this option but with others it may be a possibility when the number of reasonable incorrect options is limited.
- Use humour in moderation. McMorris et al (1997) concluded that humour is a good aspect to use in assessment. From reviewing this study it seems likely that humour is appropriate if used in small classroom assessments when students are familiar with the teacher but not for more formal testing.

(Adapted from Haladyna et al, 2002)

More recently the number of guidelines for producing multiple choice questions has been reduced. Moreno et al (2004) constructed an improved set of guidelines which avoided the repetition for which previous guidelines have been criticized. The guidelines were condensed to 12 items.

- The content should avoid trivial items.
- The representativeness should guide how simple the question is.
- The main point should be in the stem.
- The grammar must be correct, questions should not be too long or too short and negative expressions should be used carefully.
- The semantics should match the content.
- There should be only one correct answer and plausible incorrect answers.
- The position of the correct answer should be random.

- The optimum number of options is three.
- The format should be vertical.
- The options should appear structured.
- Options should be independent of each other, e.g. none of the above and all of the above should be avoided.
- No option should stand out from the rest.

(Adapted from Moreno et al, 2004)

In order to test the validity of these guidelines, a variety of people involved in test construction completed questionnaires to give their opinions. The resulting opinions emphasized that more than three options can be effective, ‘none of the above’ is occasionally appropriate and ‘the options should appear structured’ seems unnecessary. In addition to this they felt the 1st, 2nd, 5th and 10th guidelines were unclear and should be rewritten. Finally the participants commented that a significant number of the guidelines were oversimplified and suggested a reorganization including grouping some of the guidelines together. In response to these results Moreno et al improved the guidelines to a total of fifteen.

- The domain of the assessment should be clear.
- The context which the test is to be used in should be specified.
- The questions should be based on the domains of interest.
- The questions should be presented clearly with no unnecessary difficulties.
- The number of questions should be representative.
- The options should be as short as possible.
- Most of the time there should only be one correct option.
- Questions should be presented vertically.
- Each option should be independent from the others.
- Options should be presented in a suitable order.
- Incorrect options should be plausible.

- Clues should be avoided.
- No option should stand out from the rest.
- Three options are usually adequate but more may be appropriate.
- The correct options position should be varied.

(Adapted from Moreno et al, 2004)

CHAPTER 3
Feedback Questionnaires

3.1 Designing and implementing the Pilot Questionnaires

We wish to know how Statistics students actually feel about the feedback they receive. The first step towards this was surveying a small group of Statistics students at Glasgow University as a pilot. A questionnaire was constructed for this purpose, using the guiding principles of effective feedback. This questionnaire was implemented by myself to the Statistics 2S class in Session 2007-08 after they had received their first marked assignment (submitted week 7, returned week 9 of semester 1). This Level 2 module is taken by mathematical sciences students, and the assignment was a short report on an analysis of data lab. The responses were collated in the form of an Excel Spreadsheet.

For the class's next assignment, (submitted week 11, returned week 12), the way feedback was produced was changed in line with the guiding principles, after reviewing the students' comments about the first assignment. In order to investigate whether the change in feedback improved students' satisfaction, the questionnaire was administered again; it was the same as the original apart from one question being adjusted and an extra question being added to account for it being a follow up questionnaire. The follow up questionnaire was implemented with the same class but, since the survey was anonymous, the two sets of comments could not be matched up. These responses were again collated in the form of an Excel Spreadsheet. Both of these questionnaires can be found in Appendix A.

3.2 Initial Findings

Despite the two questionnaires being implemented to the same class there were more responses to the original questionnaire. This is because the attendance was better the first time around. It is quite a significant difference with the first questionnaire receiving 49 responses and the follow up only 31 responses (in a class of 56 students). With regards to previous Statistics courses, over half who completed the second questionnaire had taken Statistics 1Y/1Z (a level 1 course that is strongly recommended but not compulsory), approximately 37% had not taken a Statistics class previously and 6% had taken a different level 1 course.

Both of these assignments were marked out of 20 but the performance on the second assignment was substantially worse. In the first assignment no one received 0-5, however in the following assignment just over 3% received 0-5. With the first assignment 55% received 16-20 but under 20% received this in the second assignment. This would have been due to a change in difficulty and it is very unlikely it was related to the feedback. With the first assignment over 70% expected their mark and with the second over 70% performed worse than they did on the first. The rest of the preliminary findings are shown in Table 3.1.

Table 3.1 Preliminary Findings from the Statistics 2S Class

	Original Questionnaire n = 49	**Follow up Questionnaire n = 31**	**Difference in Percentages**
Students that felt feedback was prompt enough	40 (81.6%) 95% CI 68.0% - 91.2%	30 (96.8%) 95% CI 83.3% - 99.9%	15.2% 95% CI 2.6% - 27.6%
Felt they received the right amount of feedback	36 (73.5%) 95% CI 58.9% - 85.1%	25 (80.7%) 95% CI 62.5% - 92.5%	7.2% 95% CI -11.4% - 25.8%
Felt they received the right amount of negative feedback	35 (77.8%) 95% CI 62.9% - 88.8%	28 (90.3%) 95% CI 74.2% - 98.0%	12.5% 95% CI -3.5% - 28.5%
Felt they received the right amount of positive feedback	32 (68.1%) 95% CI 52.9% - 80.9%	22 (71%) 95% CI 52.0% - 85.8%	2.9% 95% CI -17.9% - 23.7%
Felt they received the right amount of detail	27 (60%) 95% 44.3% - 74.3%	24 (77.4%) 95% CI 58.9% - 90.4%	17.4% 95% CI -3.1% - 38%
Made changes due to the feedback	31 (67.4%) 95% CI 52.0% - 80.5%	27 (87.1%) 95% CI 70.2% - 96.4%	19.7% 95% CI 1.7% - 37.7%
Felt it was useful overall	38 (82.6%) 95% CI 68.6% - 92.2%	30 (96.8%) 95% CI 83.3% - 99.9%	14.2% 95% CI 1.6% - 26.8%

Looking at the table we can see that with the second assignment students are much happier with how promptly the feedback was returned. After the feedback was changed more students also felt they had received the right amount of feedback. The table shows higher percentages for every aspect for the follow up questionnaire.

95% confidence intervals were produced for the differences between the two proportions to test if there were significant differences between the responses to the original and follow up questionnaires. The proportion of students after the intervention who felt feedback was returned promptly enough is highly likely to be larger than the proportion of students before the intervention by somewhere between 2.6% and 27.6%. The proportions of students before and after the intervention who felt they received the right amount of feedback are not significantly different. The proportion of students after the intervention who felt they received the right amount of negative feedback is highly likely to be larger than the corresponding proportion of students before the intervention by somewhere between 2.5% and 35.2%. The proportions of students before and after the intervention who felt they received the right amount of positive feedback are not significantly different. The proportion of students after the intervention who felt they received the right amount of detail is highly likely to be larger than the proportion of students before the intervention by somewhere between 2.1% and 42.6%. The proportion of students after the intervention who made changes due to the feedback is highly likely to be larger than the proportion of students before the intervention by somewhere between 6% and 41.8%. The proportion of students after the intervention who felt the feedback was useful overall is highly likely to be larger than the proportion of students before the intervention by somewhere between 6% and 32.5%.

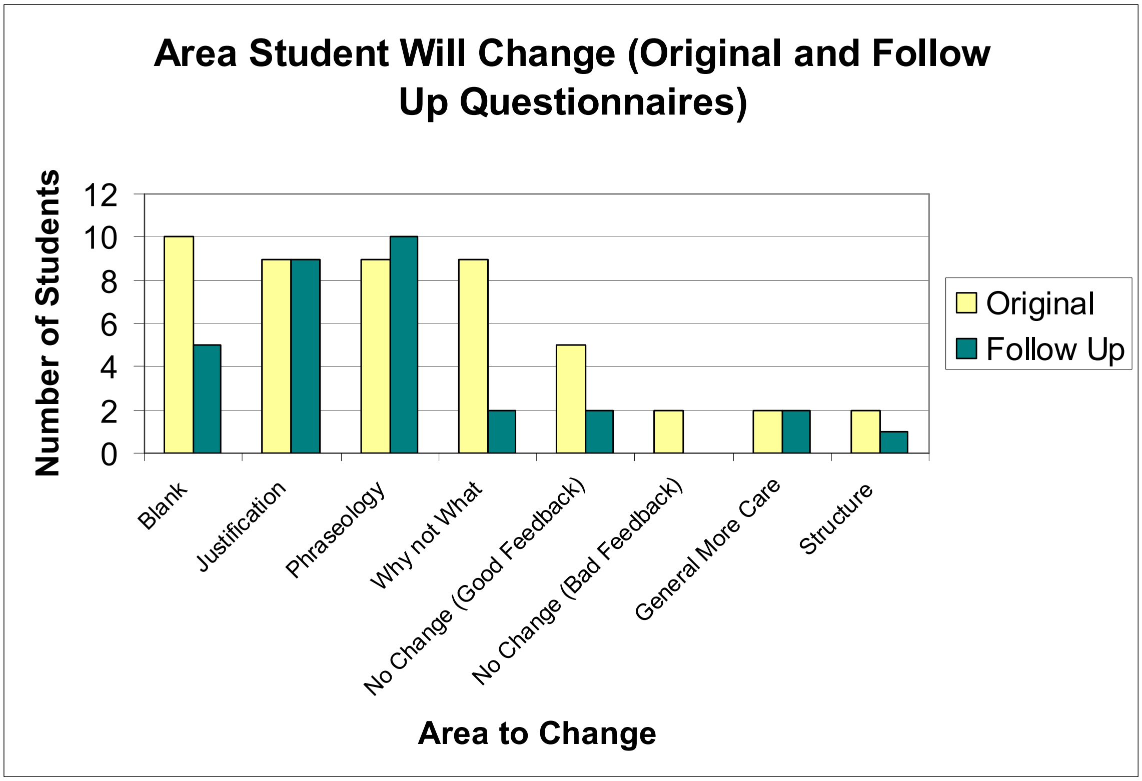
This appears to demonstrate how much more satisfied the students were after the feedback was changed in line with the guiding principles. The confidence intervals for differences in percentages are not strictly valid since some of the same respondents were included on the two occasions and, therefore, the data in the two samples are not independent. In the absence of response bias, this would simply have the effect of making the confidence intervals wider than they would have been had it been possible to match individuals' responses on the two occasions. However, there is a danger of response bias to the second questionnaire, as the students who did not submit a response were those who missed a lab and they might, therefore, be less motivated and engaged than the respondents.

20% more students stated they would change their work in response to the feedback after the feedback given was improved. This is quite a high percentage change and it is likely this is due to the feedback being more productive. It may also be due to the students making similar mistakes on both reports and it taking repeated comments for them to decide to change. This seemed to be the case for quite a lot of students. Students were asked to comment on what they would change and these statements were categorized. Table 3.2 shows what these categories represent. The results of this for both the original and the follow up questionnaire are shown in Figure 3.3.

Table 3.2 Categories

Category	**Explanation**
Blank	Student gave no answer.
General More Care	Student felt they could improve careless mistakes.
Justification	Student needs to justify their answers more.
No Change (Bad Feedback)	Student could not improve because the feedback was unhelpful.
No Change (Good Feedback)	Student could not improve because their mark was Satisfactory.
Phraseology	The Statistics phraseology they used was incorrect.
Structure	Their structure of the report could be improved.
Why not What	Student gave an explanation of why they should change their work and not what they will change.

Figure 3.1 Areas for Change



As Figure 3.1 shows, the same problems occur on both assignments. Justification and Phraseology both have a high number of students who need to improve in these areas. In the follow up questionnaire less students answered why they would change their work rather than what they would change. This could be because the students were more familiar with the questionnaire and understood more fully what was being asked. This may also explain why fewer students left this question blank on the second questionnaire.

3.3 Designing and implementing the updated questionnaires

An updated version of this questionnaire was implemented to the Statistics class, Statistics 1C, after the completion and return of their first mini project. This class is Statistics for Psychologists and Social Scientists. This project was different from the 2S report as it was worth 10% of student's total assessment whereas the 2S report was only worth 5%. They also differed on the marking of the assessments. For the 2S report they were all marked by the same individual; the class lecturer. However in the S1C project they were marked according to the students' lab group, with each lab tutor marking their own group.

The following changes were made to the questionnaire for use with this class. Their project was out of 25 so the first question had to be changed to reflect that. We decided to add a section on self esteem because the literature revealed that this can be connected to how people view feedback on their work. We used the Rosenberg Self Esteem Scale (Rosenberg, 1965) because this is the most widely accepted measure of self esteem. This was inserted into the questionnaire and not given as a separate document. We also added a question on how fair students felt their mark was, to link to the self esteem scale. We changed the question on detail from being open ended to a list of categories where students can tick as many as apply. These categories were created from comments given in the pilot questionnaires. Finally we added a question on how much attention students pay to feedback in general. This was to reveal if the questionnaire was forcing students to engage with feedback more than usual. This questionnaire was then implemented by myself during the S1C labs. The following standard statement was read before distributing the questionnaires.

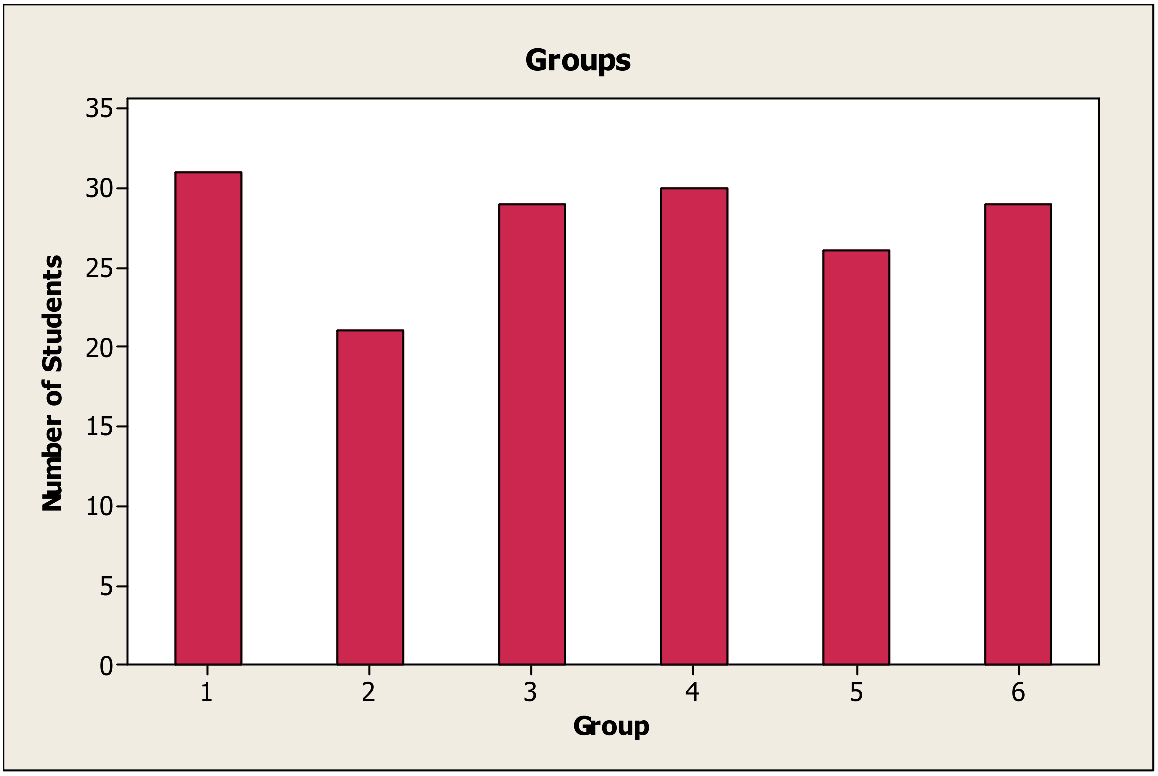
Hello everyone my name is Karina Paterson and I am a postgraduate in the Statistics department. My research is into the principles of feedback and today I am hoping you can help with this. All that is required is five to ten minutes of your time for a quick questionnaire. I am funded by the higher education academy and we are looking for your honest and personal views. Obviously all of the data will be treated anonymously but if you do not want to take part or leave out any part then that is fine. The results are going to be used to guide statistic departments in the whole of the United Kingdom so your inputs will be really useful.

The responses were collated in the form of an Excel Spreadsheet. Each question on the self esteem scale was scored using a three point Likert scale. This meant individual scores for self esteem could range from 0-30, with a higher score meaning higher self esteem.

3.4 Initial Findings

There were six different lab groups and the total number of students who responded was 166, about 70% of the class. The number of students who responded for each group is shown in Figure 3.2.

Figure 3.2 Student Response in Each S1C Group



The number of responses is similar for all six groups with the exception of group 2 having quite a low response of 21 students. The highest number of students (44.6%) achieved between 16-20 for the project. This was closely followed by the 40.4% who achieved 11-15. Only 6% received the highest marks, 21-25. These results are reflected in student's expectations. 30.1% of students achieved a mark worse than they expected, with only 18.1% attaining a result better than they expected. This may be due to some issues with the marking scheme for the project. The project involved a question on twins where the vast majority of students used a two sample t test. However the marking scheme only gave marks for a paired analysis, which many of the markers felt unfair. This resulted in very few students receiving any marks for that question. Regardless of this the majority of students felt the mark they were given was fair (95.1%). No students commented that their mark was unfairly high. Most students were also satisfied with how quickly the project was returned (90.9%). The results for amount of feedback, whether it was detailed enough and helpfulness are shown divided by lab group in Table 3.3. This was to compare the quality of feedback between different individuals.

Table 3.3 Results by Lab Group

	Group 1 n=31	Group 2 n=21	Group 3 n=29	Group 4 n=30	Group 5 n=26	Group 6 n=29
Received the right amount of feedback	22 (71%)	15 (71.4%)	22 (75.9%)	19 (63.3%)	16 (61.5%)	22 (75.9%)
Feedback was detailed enough	17 (54.8%)	11 (52.4%)	19 (65.5%)	17 (58.6%)	14 (53.9%)	17 (58.6%)
Feedback was overall helpful	28 (96.6%)	16 (76.2%)	26 (92.9%)	25 (89.7%)	23 (88.5%)	25 (89.3%)

The percentage of students satisfied with the amount of feedback is quite similar for all groups. The lowest is group five with 61.5% and the highest are groups three and six with 75.9%. A Chi-Squared test of homogeneity gave a chi-squared value of 2.731 (on 5 degrees of freedom) , p = 0.741.

This shows we cannot reject the null hypothesis that the proportion of students who felt they received the right amount of feedback is equal for all lab groups.

The results are substantially poorer for the question on whether the feedback was detailed enough. Most groups only have just over half their students satisfied with the detail. This is with the exception of group three where more students felt enough detail was given (65.5%). A Chi-Squared test of homogeneity gave a chi-squared value of 1.250 (on 5 degrees of freedom) , p = 0.940.

This shows we cannot reject the null hypothesis that the proportion of students who felt they received the right amount of detail in their feedback is equal for all lab groups.

A cross tabulation was carried out for the results on amount and detail to reveal if students were happy with one and not the other. This is shown in Table 3.4.

Table 3.4 Amount vs Detail of Feedback

Rows: Amount   Columns: Detail

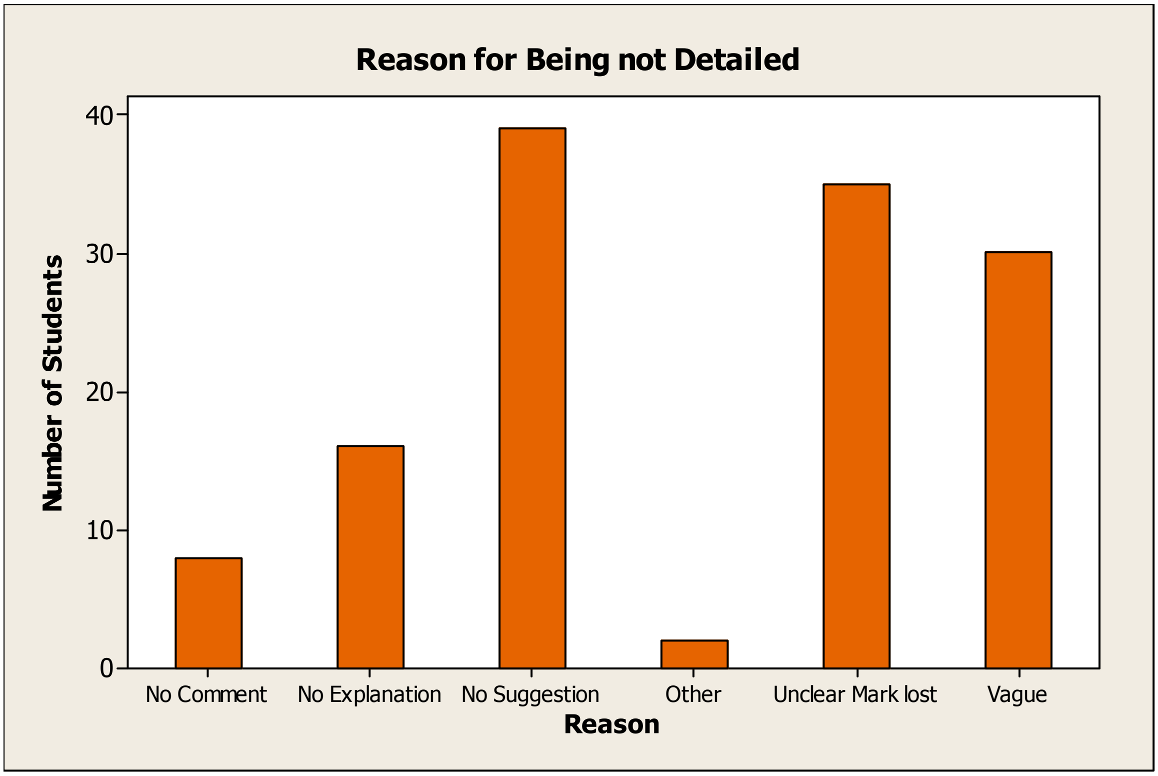
	Not Enough	Right Amount
No Too Little	43 (87.8%)	6 (12.2%)
Yes	26 (22.6%)	89 (77.4%)

A Chi-Squared test of association gave a chi-squared value of 59.834 (on 1 degrees of freedom) , p = 0.000.

This shows we can reject the null hypothesis of independence and conclude that, in this population, there is an association between how satisfied students are with the amount of feedback they receive and the detail of it. However, 26 students felt that there was the right amount of feedback but it was not detailed enough, whereas only 6 students thought there was the right amount of detail but not enough of it. It appears the detail of the feedback needs to be improved more than the amount that the markers give.

Figure 3.3 reveals the reasons students gave for why feedback was not detailed enough.

Figure 3.3 Why Feedback was not Detailed



The most common reason is that no suggestion was given for improvement. It being unclear where a mark was lost and the feedback being too vague were also frequently given as reasons.

The responses about the overall helpfulness are the most positive in Table 3.3. These figures also have the highest range with the lowest being group two at 76.2% and the highest group one at 96.6%. A Chi-Squared test of homogeneity gave a chi-squared value of 5.831 (on 5 degrees of freedom) , p = 0.323.

This shows we cannot reject the null hypothesis that the proportion of students who felt the feedback was helpful overall is equal for all lab groups.

With regards to amount of feedback students also answered for both positive and negative feedback. More students felt there was not enough positive (45%)

compared to not enough negative (19%). Very few students commented they received too much positive (0.7%) or negative (2.6%) feedback. 77% of students stated they would change their production of reports due to the feedback they received.

The majority of students believe they pay the same attention to feedback as the average student (69.9%). 16.6% felt they pay more attention than the average student and 12.3% that they pay less attention. Only 0.6% claimed they did not pick up feedback. Students self esteem scores appeared to differ with these categories. The students who pay more attention and the same attention have mean scores of 22.7 (S.D. 5.6) and 22.3 (S.D. 5.5) respectively. These are similar to the overall self esteem mean which is 21.9. However those who pay less attention have a mean self esteem score of 18.8 (S.D. 6.4). A one way ANOVA gave F = 3.74 (on 2 and 155 degrees of freedom), p = 0.026.

The p value is less than 0.05 so we can reject the null hypothesis that the mean self esteem scores are equal and conclude that at least two of the population means are different. Therefore multiple comparisons were produced using Tukey's intervals.

More - Less	4.08	(0.10, 8.06)
More – Same	0.39	(-2.50, 3.30)
Same – Less	3.69	(0.44, 6.94)

It can be concluded from the Tukey's intervals that the students who paid the same attention as the average student have a mean self esteem score that is significantly higher than that of students who pay less attention than average, and this mean difference is highly likely to be between 0.44 and 6.94 units on the self esteem rating. The students who pay more attention than the average student have a mean self esteem score that is significantly higher than that of students who pay less attention than average, and this difference is highly likely to be between 0.11 and 8.06 units on the self esteem rating. There is no significant difference in mean self esteem score between those who pay the same attention as the average student and those who pay more attention.

As pointed out in Chapter two, the literature suggests it may be lower self esteem which is causing students to pay less or no attention at all to feedback. They might be afraid of receiving negative feedback which could lower their self esteem even more. There was also an indication of a difference in self esteem between those who felt their mark was fair and those who felt it was unfairly poor. The students who felt it was unfairly poor had a higher mean self esteem (24.0) than those who felt it was fair (21.7). This is to be expected as the students who think they should have had a better mark will have a higher opinion of their work and themselves. Self esteem did not appear to have a relationship with any other questions.

The overall mean self esteem score was 21.9 with a standard deviation of 5.8. This seems quite high compared to other studies in the literature of British students also using the Rosenberg Scale. Begley and White (2003) conducted a study using students in a nursing school which included how self esteem changes during a three year preregistration programme. At the beginning of the programme the mean self esteem was 10.6 with a standard deviation of 3.5. Towards the end of the programme the mean self esteem was 9.2 with a standard deviation of 3.3. Another study using nursing students found a mean self esteem of 19.5 and a standard deviation of 4.5 for students who received structured tutorial support and a mean self esteem of 14.8 and a standard deviation of 3.4 for those who did not (Gammon and Morgan-Samuel, 2004). These means are much lower than the mean this study observed. However these students are studying a different subject and that could be an important factor. Regardless of this, a study using general undergraduate students found a mean self esteem score of 18.2 with a standard deviation of 4.5 (Pulford et al, 2005). Nethertheless the students in this study are primarily psychology students. It may be that psychology attracts people with a high self esteem. One reason for this could be how competitive the subject is, making it difficult to be a success in the field. This is unlikely to appeal to anyone with a low self esteem. The fact all the students are from Glasgow University may also play a part. It is a prestigious University and has one of the highest entry requirements in Scotland to study psychology.

3.5 Guiding Principles – Briefing Document

Another source of evidence for the quality of feedback currently given to students, reviewed by me in May 2008, is the reports and projects unclaimed by students from the Statistics class S1C in Session 2007-08. Many of these were examined. The students were identified on their reports only by enrolment number and the markers were anonymous so their comments cannot be attributed to them directly by name.

The first clear issue was that some of the assignments had no written comments on them at all. This is most worrying when the student did not receive full marks. This is unacceptable as the student has no idea where the marks were lost. More commonly no feedback is given when the student receives full marks. This fits with the literature which revealed positive feedback is given much less frequently than negative.

In the instances where positive feedback was given, the comments were largely very general.

“Very good” - Report L2, student received 9/10.

“Excellent” - Report N2, student received 10/10.

This is an improvement from no feedback but the guiding principles also focused on how unhelpful general statements are. Some of the positive feedback given did specify the area of the assignment that deserves praise.

“Good conclusions” – Report N2, student received 6/10.

However there is no mention of why this was a good conclusion. A perfect example of positive feedback is the following.

“This is an excellent plot for question 2 it shows that the two data sets are dependent” – Project 1, student received 10/25.

The tutor gives the right amount of detail but unfortunately this is a rare example. However there are more examples of detailed feedback when the comments are negative. This is essential since vague criticism is unlikely to be of any help to students. The following examples are really useful as they explain exactly what was expected of the student.

“Question 4 demands: Descriptive analysis – Scatterplot and first impression on the nature of the relationship. – Formal analysis – Regression analysis, validation of assumptions, prediction and conclusion” – Project 1, student received 10.5/25.

“One should have expected to see – Descriptive statistics of the difference between the two IQ’s, graph of the difference, statement of first impressions” – Project 1, student received 10/25.

Some tutors have even given constructive criticism when a student received full marks.

“Next time you can reduce the font size so that everything will be accommodated” – Report F3, student received 10/10.

“Safer to use the probability plot to establish normality” – Report F3, student received 10/10.

These are great examples of tutors putting in the right amount of effort. Unfortunately it may be the case that all of these examples came from the one tutor. There is also a fair amount of negative feedback that is too vague.

“Continuum response, covariate and categorical explanatory etc” – Report L3, student received 9/10.

“2 groups, continuous, in range of etc” – Report F3, student received 7/10.

It seems unlikely the student would know exactly what to change from these statements. Referring back to the guidelines, criticism without any suggestion for improvement should be kept to a minimum. Here are some examples of tutors following this rule.

“This gives a wrong representation of the populations responses. Percentages should be within levels of region” – Report N2, student received 6/10.

“This is a wrong test. Remember you are comparing two populations you should have used 2 sample t test” – Project 1, student received 10/25.

However not all of the feedback is explicit about what should be changed.

“You have treated age as a categorical variable here” – Report L3, student received 9/10.

Here it appears the student’s work is just criticized. As the literature revealed, this can be quite damaging for a student. Another issue the guidelines identified which some of the tutors have been using is phrasing negative feedback as questions.

“Randomness of samples?” – Report 4, student received 8.5/10.
“Effect of age?” – Report L2, student received 9/10.
“Validity of Chi Square?” – Report L3, student received 9/10.

Many students are confused by these questions and do not identify them as negative feedback which requires them to change something. One report’s only comment was a question even though the student lost two marks.

“Sq root transform?” – Body Fat**,** student received 8/10.

This short question does not give enough information on where the marks were lost. All of these questions could be phrased as suggestive statements such as the following.
“You must state your reasons for saying this” – Project L2, student received 8/10.

This statement is much more helpful than the frequently written “Why?”

More examples of good detail could be found in the summary comments at the end of the reports and projects.

“Almost all marks you lost here were lost at the subjective impression stage” – Project N2, student received 6/10.

Feedback at the end of an assignment serves as a useful overview of what the student should concentrate their attention on. Several of the tutors use this as an opportunity to combine negative and positive feedback, helping to soften the criticism.

“Good you were on the right track except you did not finish” – Report N2, student received 6/10.

“Good but your descriptive analysis was incomplete” – Project N2, student received 6/10.

However the guidelines do warn against the use of the word “but” and similar words as they can devalue the praise that has just been given. This could make giving negative and positive feedback in the one summary quite difficult. The following example avoids these words but it may still have the same effect.

“Good, you omitted some important comments and conclusions” - Report L2, student received 6/10.

The next example is a very valuable summary as it ends with positive comments but a reminder to study the previous constructive criticisms.

“Very good. I hope the comments above will be helpful for future reports” – Report M3, student received 8/10.

Despite the effectiveness of these summaries they are sadly infrequent in the S1C reports and projects.

The last area from the guidelines to be discussed is how feedback should be objective. One way markers can keep their feedback objective is by wording their comments as their personal opinion. This can help to mitigate criticism as well as showing the feedback may not be a collective opinion.

“I think the aim has to do with testing assumptions of normality” – Body Fat**,** student received 7.5/10.

“It appears you used the wrong set of data as your result differ from what is expected for the data” – Report L2, student received 6.5/10.

Finally there are some beneficial aspects in several of the feedback examples that have been overlooked in the guidelines. One tutor marks a zero beside the areas a student lost all the marks on. This makes it clear to the student exactly where they are losing marks, which some of the tutors failed to convey. Some of the assignments even had diagrams drawn on by the markers, to communicate more effectively how students’ graphs were incorrect. This is much clearer than written comments beside a graph. Lastly a few of the markers’ comments included underlining a significant part of the feedback.

“If they seem like they are not parallel that suggests that an interaction term **is** necessary” – Report M3, student received 8.5/10.

The underlining emphasizes where the student was confused and is good practice when writing feedback.

These findings show that in the S1C marking there is a fair amount of really good feedback being written. Many of the examples show great detail and creative thinking from the tutors. However there is also a significant number of instances where the feedback is not good enough and sometimes non existent. Although the tutors were anonymous, their handwriting was distinctive enough for it to become clear that the majority of useful feedback comes from a couple of tutors. This is a very real problem as all of the tutors will have the same time constraints and are paid exactly the same.

In the light of the literature on good feedback, reviewed in Chapter 2, the completed questionnaires discussed in Sections 3.1 and 3.3, and the quality of the feedback written on the Statistics 1C reports that were reviewed, we decided to produce a condensed version of the guiding principles of feedback that could be issued to assist markers. The completed briefing document is included as Appendix B.

CHAPTER 4
Constructing Quiz System

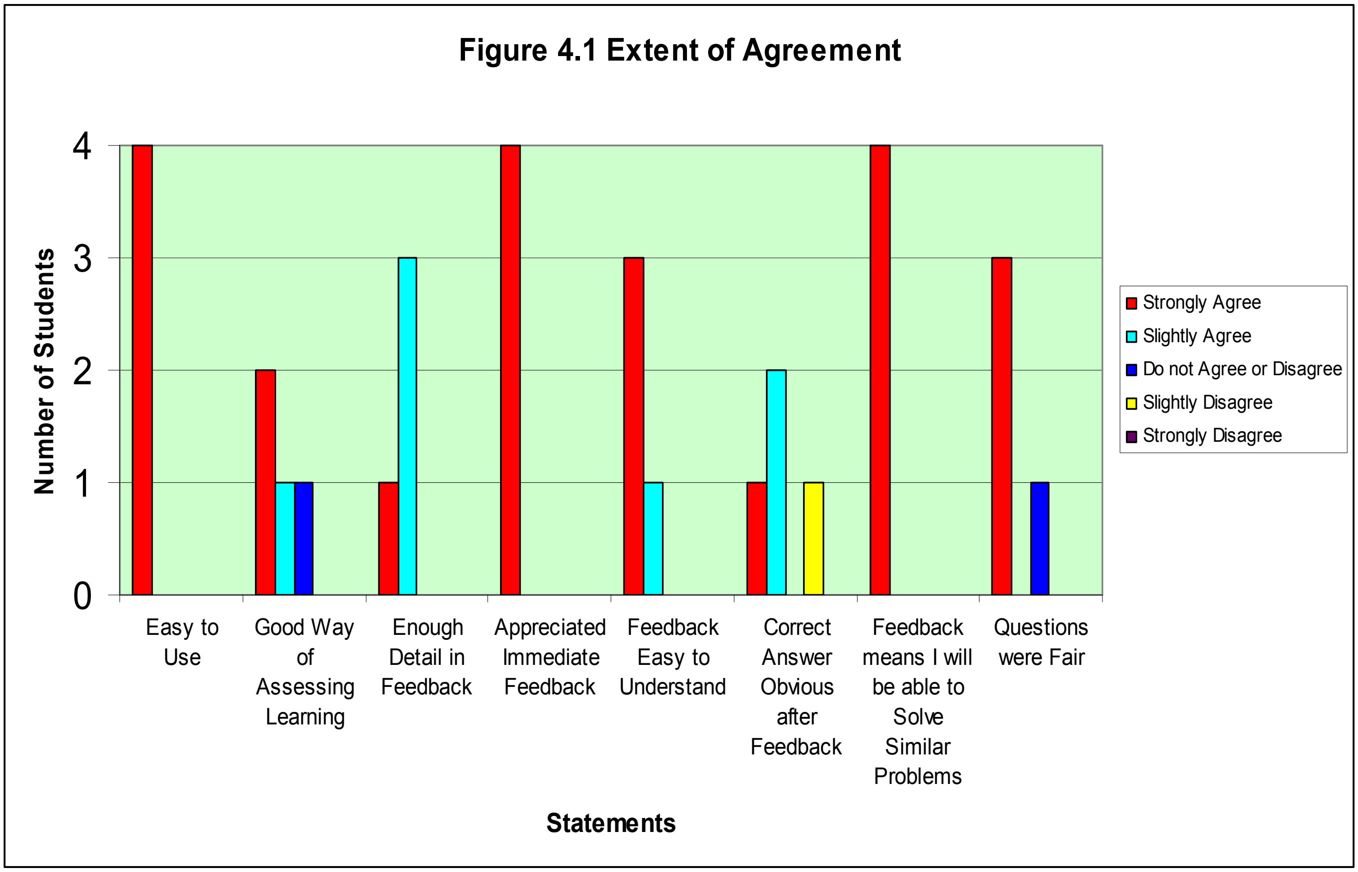
4.1 Piloting Model Choice

Model Choice is a computer based multiple choice assessment system that was created within the Department of Statistics (mainly by John McColl and Ewan Crawford, with input from Professor Helen MacGillivray, Queensland University of Technology) for initial use with its level two Statistics classes as formative assessment. The questions are on probability models and there are three difficulty levels, basic, intermediate and advanced with each level containing 100 questions. Students are given ten questions, chosen at random from a large pool of questions at the chosen difficulty level, and are allowed two attempts to identify the correct answer. The student system may be accessed at the following web page: http://www.mathstore.gla.ac.uk/modelchoice/ using the following log-in information, Name: guest and Password: guest.

The defining feature of model choice is the immediate feedback that students are given. For incorrect answers the feedback tries to lead the students to the correct answer on their second attempt and for correct answers an explanation is also given. There is also a '"do not know"' option for the first two levels and the feedback for this is essentially a clue to the correct answer. At the end of the test the student can see how many they answered correctly the first time and how many after a second attempt.

A small pilot was conducted with four Statistics students. These students worked through Model Choice at the basic level and then completed a questionnaire on how they found this. Two of the students received 6-8 out of 10 and the other 2 received 9-10. For two of the students this was as expected, one student felt it was worse than expected and one that it was better than they expected. Figure

4.1 shows the extent to which the students agreed or disagreed with a collection of statements.



The only statement there is any disagreement on is ‘The correct answer was obvious after the initial feedback’. One student slightly disagreed and the other three agreed to some extent. This is not necessarily a problem because the answer is not intended to be too obvious after the feedback. If the answer was too obvious students would not be testing themselves on their second attempt.

The most frequent response to all statements was strongly agree meaning the program received a very positive reaction. This was apart from ‘There was enough detail in the feedback’ where three of the four students only slightly agreed.

To receive more detail from the students they were asked open ended questions at the end of the questionnaire. The first question was “What did you like best about the program?” Three of the four students emphasised how easy the program was to use. One student felt this would encourage more students to revise since it is easier than paper and pencil revision. Students also made the following comments.

“I thought the questions were a really good length, any longer and I would have forgotten the point.”

“It is a great method of revision.”

“It gives you a wake up call, it shows that you “do not know” as much as you think you do.”

“I liked how the questions used real life situations.”

“The questions were very detailed which is helpful.”

The students were also asked what they did not like about the program. All of the students struggled to think of something negative about the program. Two students said there was nothing bad about the program. After some thought the other students made the following comments.

“The “do not know” option, students may be tempted to just put that.” This student may have misunderstood the purpose of the “do not know” option. If the program is being used as an assessment measure students will not receive the full marks for the question meaning the “do not know” option is not an easy way out. On the other hand if the program is being used for revision then this option will assist them in finding the correct option and therefore help them learn.

“It is demotivating to get an answer wrong” This is a valid point but unfortunately it is an essential part of learning. The student did end this comment on a positive note. “Only for some people though, for others it will show them they need to revise more.”

4.2 Constructing S1C Questions

After piloting model choice we decided a similar program would be useful for giving students immediate feedback at the end of some of the Statistics S1C labs. Before designing the program we decided to take some of Gal and Garfield’s advice on board (see Chapter one). The system will use real stories and data sets and multiple questions will be asked on the same data. Some questions will also contain more information than is necessary so students have to be selective.

The Lab Co-ordinator, Dr. Mitchum Bock, indicated four labs where he felt it would be constructive to implement computer based assessment paired with immediate feedback. These labs were Lab E - Sampling and interval estimation, Lab J - Multiple Regression, Lab K - Experimental Design and Lab O - Categorical Data. These labs were mainly chosen because they cover important topics and the students were not required to complete a lab report on these topics. Questions were constructed based on both the context of the lab and the relevant lectures. These can be found at

http://www.mathstore.ac.uk/teststat/intranet/	staff site
http://www.mathstore.ac.uk/teststat/	student site

The student site may be accessed using the same log-in details as on P61 for model choice.

All of the labs consist of ten types of question. In the first instance, it was decided to aim to create three or four alternatives for each question type using different data sets or theoretical questions. The data were taken from previous S1C labs, Minitab example sets and the data and story library (http://lib.stat.cmu.edu/DASL/) Each question has four options, as well as a “do not know” option, with only one being the correct answer.

Feedback was also created for each of these options to assist the students as they are given two attempts. If the student answers incorrectly they are given feedback on why that answer is false which should lead them to the correct answer. If the student answers correctly they are also given feedback on why this was correct.

The program is very similar to Model Choice but there were some aspects that had to be changed. In Model Choice the order of the options is always random but with the S1C quiz it made sense for some of the options to be in a specific order. This was mainly due to the length of the options which meant setting them in an increasing length order.

With the S1C quiz it was also appropriate to have certain questions linked together which was not a feature of Model Choice. This was so that a number of questions could use the same data set and background. Having them linked one after the other will reduce reading time for the student.

Furthermore the S1C quiz brought many different formats of questions to the program. The questions in Model Choice only ever involved text whereas the S1C quiz uses graphs, tables and Minitab output to illustrate the concept.

In Model Choice all of the questions were on the same topic but for S1C all of the question types had to be named on the computer system. Each time a student is tested, test items are selected by stratified random sampling, one from each question type. The question types for each lab are shown below.

Lab E - Sampling Distributions

Type 1 - Point Estimates

Type 2 – CI for one mean - interpretation

Type 3 – Using MINITAB - trials and events

Type 4 – CI for one proportion - interpretation

Type 5 – Miscellaneous knowledge

Type 6 – Matching CI with sample size

Type 7 – Spotting the wrong CI

Type 8 – Two samples - boxplot

Type 9 – CI for two means - interpretation

Type 10 – CI for two proportions - interpretation

Lab J – Multiple Regression

Type 1 - Meaning of slope parameter in simple regression.

Type 2 - Interpreting residual plots.

Type 3 - Identifying the predictor and response variables.

Type 4 - Interpreting matrix plots.

Type 5 - Interpreting correlation matrices.

Type 6 - Interpreting regression output.

Type 7 - Assumption checking theory.

Type 8 - Regression theory.

Type 9 - Interpreting stepwise regression output.

Type 10 - Interpreting confidence intervals and prediction intervals.

Lab K – Experimental Design

Type 1 – Knowledge – Basic Design

Type 2 – Ordering

Type 3 – Knowledge – Randomising

Type 4 – One Way ANOVA

Type 5 – Knowledge – Latin Square and Design

Type 6 – Confidence Intervals

Type 7 – Confidence Intervals – Summarising

Type 8 – Pairwise Comparisons

Type 9 – Factorial Design

Type 10 – Interaction Plot

Lab O – Categorical Data

Type 1 – Dealing with Categorical Data in Minitab

Type 2 – Testing for Association – Data Description

Type 3 – Comparing Proportions

Type 4 – Validity of Chi-Square Test

Type 5 – Chi Square Analysis

Type 6 - Knowledge - Marginal Homogeneity and Chi-Square

Type 7 - Cramer's Measure of Association

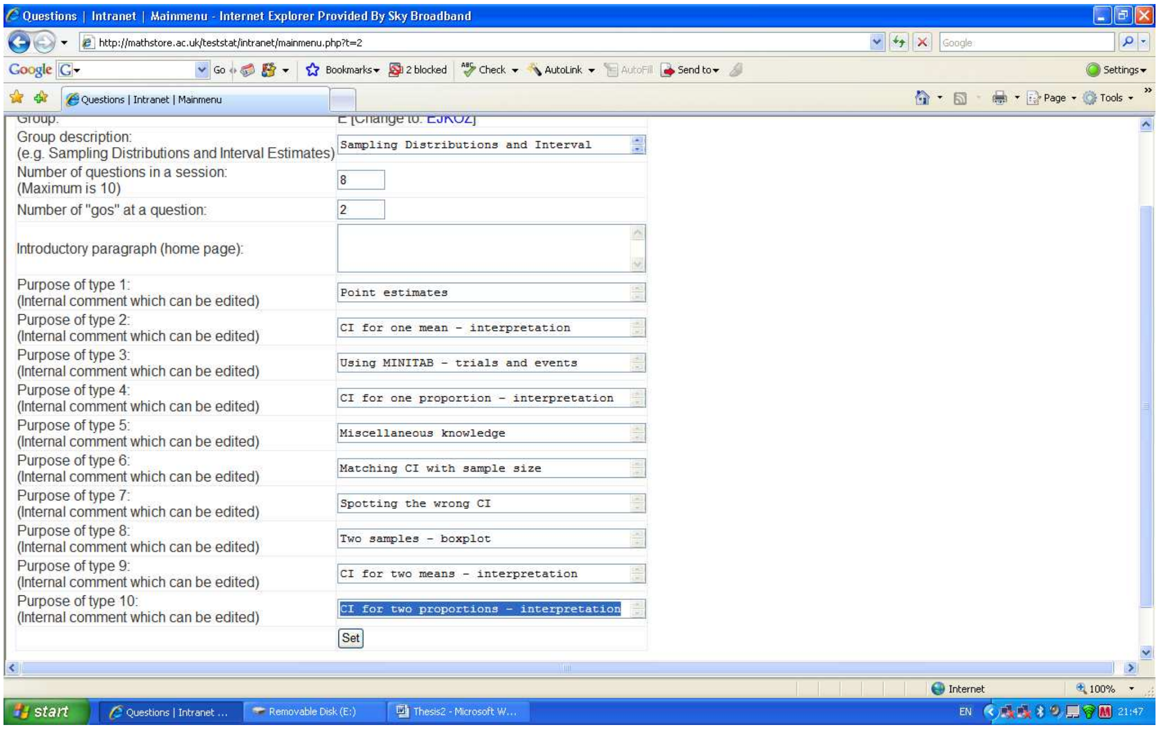
Type 8 - Test of Marginal Homogeneity

Type 9 - Bonferroni Intervals

Type 10 - Knowledge – Association and Homogeneity

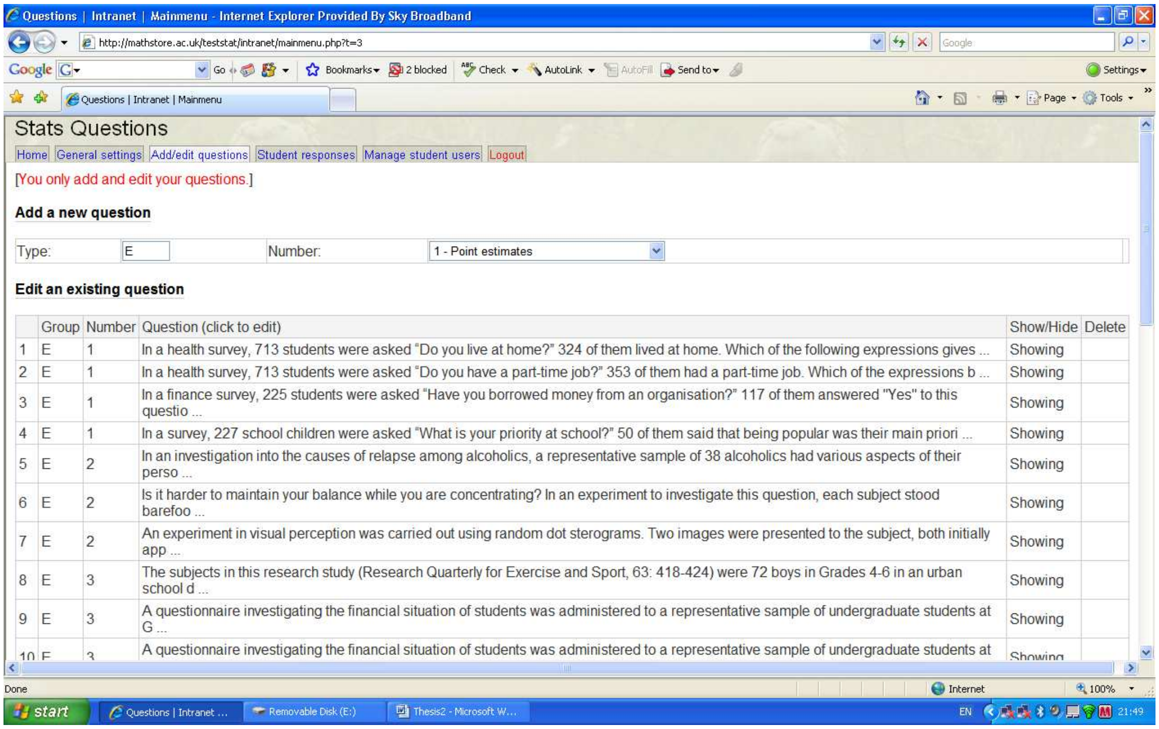
Having clearly labelled question types will help when other members of staff want to add questions to the live version of the programme. Figure 4.2 shows an example of the opening screen for tutors where the question types are shown. Figure 4.3 shows the screen where questions are added and removed.

Figure 4.2 Screenshot Opening Screen


Questions | Intranet | Mainmenu - Internet Explorer Provided By Sky Broadband
http://mathstore.ac.uk/teststat/intranet/mainmenu.php?t=2
Google
Go
Bookmarks
2 blocked
Check
AutoLink
AutoFill
Send to
Settings
Questions | Intranet | Mainmenu
Page
Tools
Internet
100%
start
Questions | Intranet ...
Removable Disk (E:)
Thesis2 - Microsoft W...
EN
21:47

Group:	E [Change to: EJKOZ]
Group description: (e.g. Sampling Distributions and Interval Estimates)	Sampling Distributions and Interval
Number of questions in a session: (Maximum is 10)	8
Number of "gos" at a question:	2
Introductory paragraph (home page):	
Purpose of type 1: (Internal comment which can be edited)	Point estimates
Purpose of type 2: (Internal comment which can be edited)	CI for one mean - interpretation
Purpose of type 3: (Internal comment which can be edited)	Using MINITAB - trials and events
Purpose of type 4: (Internal comment which can be edited)	CI for one proportion - interpretation
Purpose of type 5: (Internal comment which can be edited)	Miscellaneous knowledge
Purpose of type 6: (Internal comment which can be edited)	Matching CI with sample size
Purpose of type 7: (Internal comment which can be edited)	Spotting the wrong CI
Purpose of type 8: (Internal comment which can be edited)	Two samples - boxplot
Purpose of type 9: (Internal comment which can be edited)	CI for two means - interpretation
Purpose of type 10: (Internal comment which can be edited)	CI for two proportions - interpretation
	Set

Figure 4.3 Screenshot Adding Questions



Stats Questions

Home | General settings | Add/edit questions | Student responses | Manage student users | Logout

[You only add and edit your questions.]

Add a new question

Type: E   Number: 1 - Point estimates

Edit an existing question

	Group	Number	Question (click to edit)	Show/Hide	Delete
1	E	1	In a health survey, 713 students were asked "Do you live at home?" 324 of them lived at home. Which of the following expressions gives ...	Showing	
2	E	1	In a health survey, 713 students were asked "Do you have a part-time job?" 353 of them had a part-time job. Which of the expressions b ...	Showing	
3	E	1	In a finance survey, 225 students were asked "Have you borrowed money from an organisation?" 117 of them answered "Yes" to this questio ...	Showing	
4	E	1	In a survey, 227 school children were asked "What is your priority at school?" 50 of them said that being popular was their main priori ...	Showing	
5	E	2	In an investigation into the causes of relapse among alcoholics, a representative sample of 38 alcoholics had various aspects of their perso ...	Showing	
6	E	2	Is it harder to maintain your balance while you are concentrating? In an experiment to investigate this question, each subject stood barefoo ...	Showing	
7	E	2	An experiment in visual perception was carried out using random dot sterograms. Two images were presented to the subject, both initially app ...	Showing	
8	E	3	The subjects in this research study (Research Quarterly for Exercise and Sport, 63: 418-424) were 72 boys in Grades 4-6 in an urban school d ...	Showing	
9	E	3	A questionnaire investigating the financial situation of students was administered to a representative sample of undergraduate students at G ...	Showing	
10	E	3	A questionnaire investigating the financial situation of students was administered to a representative sample of undergraduate students at	Showing	

Once all of the questions were drafted, Dr. Mitchum Bock gave his thoughts on the two labs he is involved with (Experimental Design and Sampling Distributions and Interval Estimation). This revealed the structure of the S1C course was changing. The Sampling Distributions and Interval Estimation lab will now be split into two labs. This meant more questions had to be constructed for the Interval Estimation lab. Mitchum made significant improvements to the Interval Estimation lab. These were mostly on detailed wording of the questions and reducing some of the feedback to ensure the answer was not obvious. The questions were further reduced in length after being checked by Professor John McColl. This was to ensure they were as straightforward and clear as possible.

When constructing the questions the guidelines discussed in the literature review were referred to. One of the most repeated principles is that each question refers to specific principles in the course. This was followed closely as each question was based on a key area from the labs and the lectures. For example for the Categorical Data Lab, out of the ten types of question, seven were based on principles in the lab (Dealing with Categorical Data in Minitab, Comparing Proportions, Validity of Chi-Square Test, Chi-Square Analysis, Cramer's Measure of Association, Test of Marginal Homogeneity and Bonferroni Intervals) and three on principles from the lectures (Testing for Association – Data Description, Knowledge - Marginal Homogeneity and Chi-Square and Knowledge – Association and Homogeneity).

The guidelines also advise using a mixture of types of questions which can be shown from the ten different types which were included in each lab i.e. problem solving and knowledge. Here is an example of a problem solving question from the Categorical Data Lab.

In a study of general health, a sample of young people with a history of delinquency was compared with a sample of controls (i.e. young people with no history of delinquency). The table below presents data on all the young people in both groups who were found to have defective vision, only some of whom wore spectacles.

```
Rows: WearSpecs   Columns: Population

       Control  Delinquents     All

No           2            8      10
         4.375        5.625  10.000

Yes          5            1       6
         2.625        3.375   6.000

All          7            9      16
         7.000        9.000  16.000

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 6.112, DF = 1, P-Value = 0.013
```

Is this Chi-Squared Test valid?

A. Yes, since the expected frequencies are all greater than two.

B. Yes, since the observed frequencies are all greater than two.

C. No, since some of the observed frequencies are less than

Below is an example of a knowledge type from the same lab.

What type of analysis would you carry out to examine evidence for an association between the smoking behaviour of respondents (Smoker, Non-Smoker) and that of their parents (Both Smoked, One Smoked, Neither Smoked)?

A. Test of Marginal Homogeneity
B. Cross Tabulation and Chi Square
C. General Linear Model
D. Two Sample t Test

In creating the options, incorrect answers were based on the frequent mistakes that students make in the labs. This should mean that all of the incorrect options will be plausible to students who have failed to understand the concept. For example, in a question from the Categorical Data Lab on Cramer's measure of association, one of the incorrect answers incorporates the false idea that the value shows the direction of the association. Another example of this guideline from the Interval Estimation lab occurs in the questions on confidence intervals for two proportions. Here the incorrect answers will be chosen by students who do not understand how to change the interval into percentages or those who do not recognise what negative or positive numbers mean for the interval.

The options 'none of the above' and 'all of the above' were avoided in all questions. When reviewing the different guidelines there were two principles that could be contradictory. These are 'The position of the correct answer should be random' and 'Options should be presented in a suitable order'. For the S1C questions we decided that for some questions it would be necessary to have the options in an order to prevent the student from becoming confused. For example with this question from the Multiple Regression Lab it reads much easier if the options are in increasing length.

> A sample of 86 first year male Statistics students had their grip strength measured using a grip dynamometer. In addition, a number of physiological measurements were made on their dominant hand and arm: the width and length of the hand and the circumference of the forearm. The objective was to explore what contributes to a strong grip. Which variables should be used as predictors when fitting a multiple regression model?
>
> A. Grip Strength
> B. Hand Width, Hand Length
> C. Hand Width, Hand Length, Forearm Circumference
> D. Hand Width, Hand Length, Forearm Circumference, Grip Strength

However for the majority of the questions the computer program will randomise the order of the options. This will mean that the correct answer will be varied over the different positions, which is good practice. The guidelines also recommend that when creating a test, novel material is used and for the majority of the questions data sets have been used that students in this class have not seen before.

The guidelines do advise keeping questions independent from each other which has not been followed for the S1C test. This is because some of the questions use the same data and we decided that these questions should be linked together to minimise reading for the student. For example, in the Experimental Design Lab there is a question type involving analysis of an ANOVA and a type on ANOVA conclusions, so it makes sense for students to have the same context for both of these types. This does not appear to be a problem, since reducing information intake for the student is also an important guideline. All of the questions have the main idea in the stem and not the options, which is another essential guideline.

The most discussed guideline refers to how many options each question should have. The S1C questions have four options and a '"do not know" option'. The consensus seems to be that three options are enough but if more plausible incorrect answers can be created then this is useful.

Taking all of the guidelines into consideration the S1C questions adhere to the majority of these.

Unfortunately none of the guidelines give any advice on how to produce feedback for multiple choice tests. This is most likely because this type of assessment is essentially treated as a method of summative assessment. However it can be a very useful technique for formative assessment. When multiple choice tests are used as a learning tool, offering feedback should be a key feature. Multiple choice testing without feedback only informs students what questions they answered incorrectly. This means there could be an unnoticed common factor to these questions which the students do not understand. Feedback can highlight this factor allowing the students to progress.

Creating the feedback for these quizzes was a difficult process as a balance had to be achieved between giving advice that was helpful but not making the answer obvious. For the correct answers the feedback explains why the answer was correct in as much detail as possible. This is useful because it reinforces the important principle behind the question. It is also necessary for the students who think they know the answer but are not sure of the reasoning behind it. Creating the feedback for the '"do not know"' option was the most difficult task. With incorrect answers, where the thinking went wrong can be explained. However if a student responds '"do not know"' it is unclear exactly where they are confused. Most of this feedback was written as a general clue to the answer.

CHAPTER 5
Piloting the Statistics S1C Quizzes

Once two of the quizzes were added to the program it was time for an initial testing of the system. The two labs that were completed first were Lab J – Multiple Regression and Lab O – Categorical Data 2. All of the staff and postgraduates in the department were invited to an hour long testing session. From these eight members of staff and four postgraduates were able to attend. Before the testing session Matina Rassias, a teaching fellow, gave some feedback on the system. Her main points were the following:

- In the longer term there will need to be a bigger pool of questions.
- With the feedback there should be punctuation after 'this is correct' or 'this is not correct' before the actual feedback begins.
- Some of the feedback is too obvious and gives the answer.
- Some other pieces of feedback are not detailed enough.

In response to this feedback some changes were made before the testing session. Dashes were inserted between whether the question was correct or not and the explanation in the feedback to separate these. Some of the most obvious feedback was reduced to make the correct answer less apparent. The testing session used the questionnaire from Model Choice with a few changes. The questions on mark achieved and expectations were removed since the participants were not students. As well as the questionnaire, participants were given a blank sheet of paper to note down any Statistical problems they found. Participants were instructed to work through Practical O, then fill out the questionnaire and finally try Practical J. After this an informal group discussion took place.

Everyone agreed the system was very easy to use.

One member of staff suggested that the ‘“do not know”’ option could advise students to look at certain pages of their notes, making it more interactive. However this could be a problem if the notes or the order of the notes were changed meaning the whole system would have to be reviewed and edited in detail every session.

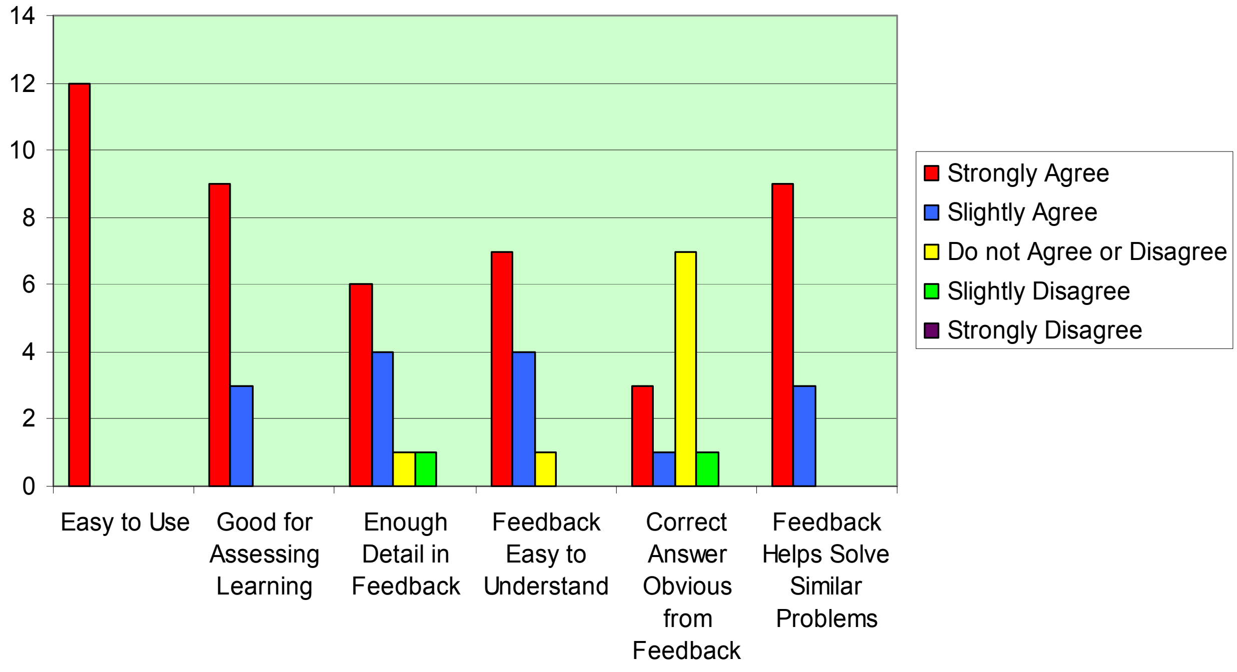
Using the system as a summative assessment method was also discussed. One possible way a score could be calculated is out of 20 marks. Two marks would be given for a correct answer on the first attempt and one mark for a correct answer on the second attempt. One issue with this is that students may avoid the ‘“do not know”’ option altogether since it counts as an attempt and they would receive feedback from an incorrect answer anyway. Guessing at an answer gives a student a one in four chance of answering correctly but ‘“do not know”’ gives them no chance. In order to avoid this the “do not know” feedback would have to be much more useful than the rest and this would have to be clear to the students. On the other hand it may be that ‘“do not know”’ would be removed for summative assessment and only retained for formative assessment. Another suggestion was that the ‘“do not know”’ option is kept separate from the options and how often students use it could be factored independently. However students would also need to be made aware of this and may still avoid it if it affects their mark.

The next comment made was that all of the questions for Lab O were quite hard with no easy questions. This was because the literature showed there was no benefit of including easy questions to build up a student’s confidence.

It was also suggested there could be a time limit for students to complete the test. This may be a useful idea since the testing session showed that participants took longer than expected, which might cause problems when running the S1C lab sessions (timed at two hours in total).

The results from the questionnaires were also largely positive. Figure 5.1 shows these results.

Figure 5.1 S1C Quiz Trial



The first question, on how easy the program was to use, received the most positive result. All twelve participants strongly agreed with this statement. Strongly agree was the most frequent response on all questions apart from whether the answer was obvious after the feedback. Most participants felt they could not agree or disagree with that statement. This is a very encouraging result since the optimum feedback would be not too obvious but give helpful advice towards the answer. Some of the participants even commented on their questionnaire that disagreeing with that statement was not a bad result. Apart from this question the question with the most variable result is related to the detail in the feedback. This was also evident from the written comments where some felt the feedback contained plenty of detail and others that there was not enough. Overall the results were very positive but it is hard to conclude a great deal with such a small sample size.

Many of the responses to the first questions were reinforced in the open ended questions. For what the participants liked best about the program five of them again highlighted how easy the program is to use. A couple of participants also remarked on how beneficial and clear the layout was. Most of the responses to this question were linked to the principles for creating multiple choice assessment which shows how carefully these were followed. For example the following statements relate to one of the most important principles.

“The wrong answers were very convincing.”

“In general the options were sufficiently challenging.”

It is essential the incorrect options are believable to students in order to really measure learning. The second comment also emphasises that the options were not too difficult. The responses also show evidence that the guideline that commonly misunderstood concepts should be assessed was met. One of the participants’ favourite aspects was that some of the incorrect options were essentially correct apart from the detailed wording. Accurate wording is very important in Statistics and this should help make this clear to students. One other participant picked up on the difficulty of providing helpful feedback without giving the answer away. They felt we had achieved this balance as much as is possible. On a similar note another participant liked how the feedback directed you back to the output. This meant students were really working towards the correct answer by themselves. Looking at what participants liked least the idea of the program being linked to more information was again commented on. Although it is not practical to make a link to notes it should be possible to link to certain web pages. Another common criticism was that some of the feedback needs to be more detailed. For example,

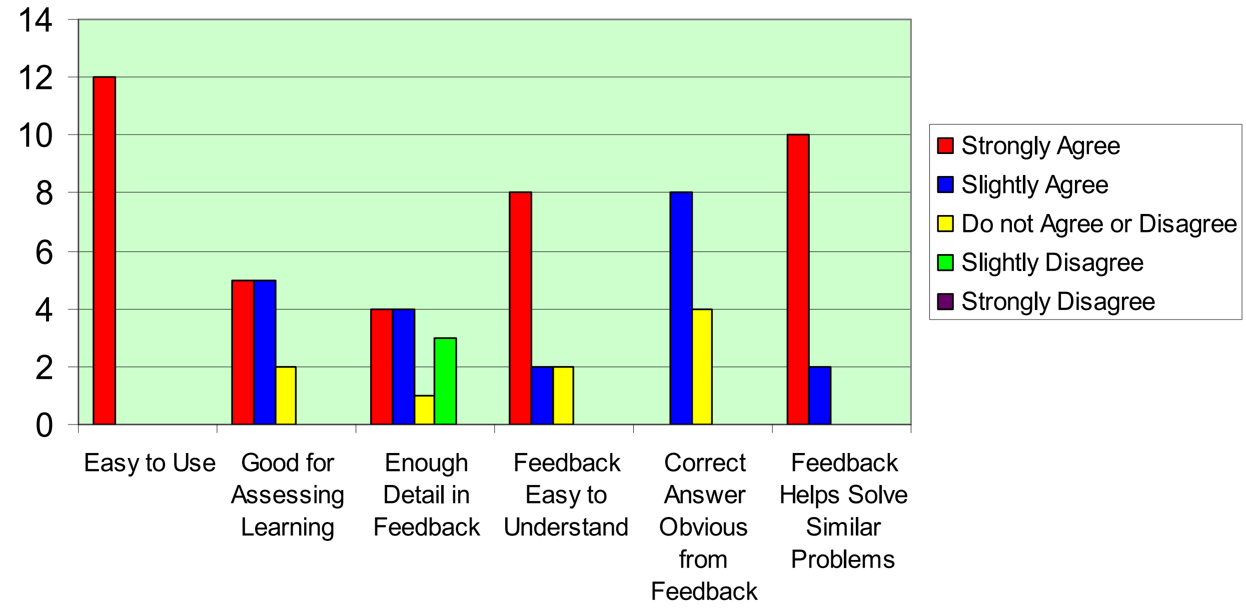
“Some of the feedback on some questions could be more explanatory.”

This is interesting because to begin with the feedback for these questions contained detail on the specific test as well as the hypothesis result but it was reduced because it gave the answer away. Again this is the difficulty of achieving the right balance of detail. A couple of participants also commented that how to reach the correct answer is not explained if the student answers both attempts incorrectly. This is a really useful observation as why the correct answer is correct is only explained if the student chooses it on one of their attempts. It would actually be more help for a student to know why the correct answer was correct if he or she had not chosen the correct answer at either attempt. Therefore we decided to always show this explanation if the correct answer had to be given to the student in such circumstances. Finally there were some small concerns over visual impact and making that stronger. Three members of staff also made written comments on statistical problems. These were mainly wording and ordering issues. One important issue this revealed is not all staff use the exact same statistical rules when teaching. The actual rule that will be taught to S1C would have to be agreed on and then used in the questions.

Another trial of the system was conducted at the beginning of the new student term. Volunteers were recruited from the third year Statistics induction lecture. Twelve students attended the half an hour trial using the same questionnaire as previously. This session was observed by Professor Helen MacGillivray (Queensland University of Technology), who had previously been involved in the development of Model Choice and was visiting Glasgow at the time.

Students were instructed to only try Practical J as Practical O is not really part of their course. Figure 5.2 shows the results from the questionnaires.

Figure 5.2 S1C Quiz Trial Students



Again all participants strongly agreed the system was easy to use. Unlike the staff and postgraduates results, strongly agree was not the most frequent response. The questions on the program being a good way of assessing learning and whether the feedback was detailed enough had a more variable result. Another difference relates to whether the feedback makes the answer obvious. No student strongly agreed with this and the majority slightly agreed. This result was not as variable as the staff and postgraduates. In relation to the open ended questions five of the students again reinforced how easy the program was to use and understand.

“The structure is easy to understand and use.”

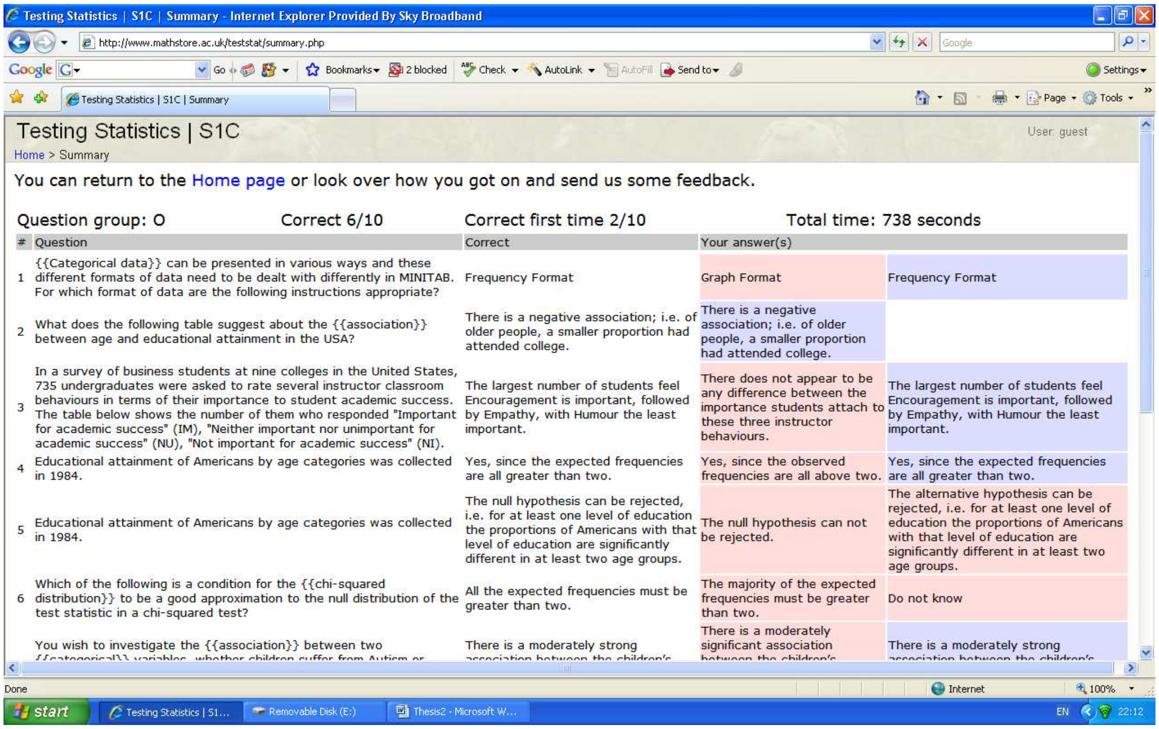
The students picked up on how advantageous the program would be as a revision exercise which none of the staff or postgraduates commented on. On a similar note some participants felt the best feature of the program was how it highlights your strengths and weaknesses. The majority of the students favourite aspect was the feedback itself.

“Getting a second chance to answer – good but I didn’t think the feedback made the answer obvious. I still needed time to think about why I was wrong”

Similar to this student many of the participants liked the second chance feature. This may be because from a student’s perspective having two attempts would make it easier. However if the system is used for summative assessment each attempt would be worth different marks. Finally one student liked how the quiz did not involve any writing. The most commonly given response for what the participants liked least was that the questions could have been spaced out better which would make it easier to read. The difficulty with this is the majority of the questions and options are reasonably long meaning spacing them out more would run onto another screen. Discussing this with the students revealed they would rather the font size was reduced and the text could be spaced out more while keeping it all on one screen.

Another frequent suggestion was that it would be useful to have an overall summary of your performance at the end of the quiz. The quiz currently informs the student how many they answered correctly on both their first and second attempts. Some of the students felt it would be better to know exactly where they had went wrong. After a discussion with Helen MacGillivray it was decided to create a summary page that showed the outcome of each question type. A screenshot of this is shown in Figure 5.3.

Figure 5.3 Screenshot of Summary Page



Testing Statistics | S1C | Summary - Internet Explorer Provided By Sky Broadband
http://www.mathstore.ac.uk/teststat/summary.php
Google
Go
Bookmarks
2 blocked
Check
AutoLink
AutoFill
Send to
Settings
Testing Statistics | S1C | Summary
Page
Tools
Done
Internet
100%
start
Testing Statistics | S1...
Removable Disk (E:)
Thesis2 - Microsoft W...
EN
22:12

Testing Statistics | S1C

User: guest

Home > Summary

You can return to the Home page or look over how you got on and send us some feedback.

Question group: O Correct 6/10 Correct first time 2/10 Total time: 738 seconds

#	Question	Correct	Your answer(s)	
1	{{Categorical data}} can be presented in various ways and these different formats of data need to be dealt with differently in MINITAB. For which format of data are the following instructions appropriate?	Frequency Format	Graph Format	Frequency Format
2	What does the following table suggest about the {{association}} between age and educational attainment in the USA?	There is a negative association; i.e. of older people, a smaller proportion had attended college.	There is a negative association; i.e. of older people, a smaller proportion had attended college.	
3	In a survey of business students at nine colleges in the United States, 735 undergraduates were asked to rate several instructor classroom behaviours in terms of their importance to student academic success. The table below shows the number of them who responded "Important for academic success" (IM), "Neither important nor unimportant for academic success" (NU), "Not important for academic success" (NI).	The largest number of students feel Encouragement is important, followed by Empathy, with Humour the least important.	There does not appear to be any difference between the importance students attach to these three instructor behaviours.	The largest number of students feel Encouragement is important, followed by Empathy, with Humour the least important.
4	Educational attainment of Americans by age categories was collected in 1984.	Yes, since the expected frequencies are all greater than two.	Yes, since the observed frequencies are all above two.	Yes, since the expected frequencies are all greater than two.
5	Educational attainment of Americans by age categories was collected in 1984.	The null hypothesis can be rejected, i.e. for at least one level of education the proportions of Americans with that level of education are significantly different in at least two age groups.	The null hypothesis can not be rejected.	The alternative hypothesis can be rejected, i.e. for at least one level of education the proportions of Americans with that level of education are significantly different in at least two age groups.
6	Which of the following is a condition for the {{chi-squared distribution}} to be a good approximation to the null distribution of the test statistic in a chi-squared test?	All the expected frequencies must be greater than two.	The majority of the expected frequencies must be greater than two.	Do not know
	You wish to investigate the {{association}} between two {{categorical}} variables, whether children suffer from Autism or	There is a moderately strong association between the children's	There is a moderately significant association between the children's	There is a moderately strong association between the children's

Similar to the staff and postgraduates trial a couple of students would have preferred the feedback to be more detailed. One of the students also disliked the program as a summative assessment as they find the lab quite a distracting environment.

CHAPTER 6
Discussion

6.1 Guidelines for good feedback

The aim of this study was to develop a set of useable guidelines for producing effective feedback and an online system for providing formative assessment and feedback to students studying Statistics as a service course. This project was quite unusual for the Statistics department since it focused on improving the actual teaching of Statistics. This meant the project produced very practical results. The areas that were focused on for improvement were giving students more effective feedback, and enhancing formative assessment and feedback through new multiple choice testing. Feedback was a sensible choice of focus since the National Student Survey (NSS) shows it to be the lowest scoring category for the majority of institutions.

Firstly the literature on assessment and feedback was reviewed. Feedback was found to have an inconsistent effect on functioning. Feedback can significantly improve performance when its purpose is clearly understood. This is often due to an increase in motivation but the feedback must be continuous. However it was shown feedback can also impair performance. This might be related to where attention is focused. When attention is directed away from the task or towards the self, performance is harmed.

There are many ideas on how to produce the best feedback. Unfortunately none of these are from the Statistics literature. However the principles that are regularly repeated should be generalisable to other subjects. These common guidelines formed the guiding principles of feedback for this project. They are: (1) balancing positive and negative feedback correctly, (2) giving the right amount of detail, (3) an appropriate quantity, (4) being objective, (5) timely and (6) future oriented.

Research was also conducted into how students use feedback. Evidence was found that students use feedback for motivation, to enhance learning, to develop reflection and for clarification. Next it was discovered there should be guidelines for students on receiving feedback. The main ideas are being able to clarify the feedback, accepting constructive criticism, approaching teachers for feedback and responding to feedback.

Finally the literature on computer assessment was evaluated. Computer assessment was found to have a lot of benefits but feedback was rarely mentioned in this context. With regards to multiple choice testing there are a collection of principles that should be followed when constructing questions. These include: using a mixture of types of questions, randomising the position of the correct answer, using new material, making sure incorrect answers are plausible and using simple vocabulary. Unfortunately no advice was found on how to give feedback with multiple choice testing.

The first pilot study was conducted with a small group of second year Statistics students. They were given a questionnaire before and after the feedback had been changed in line with the guiding principles. The questions that illustrated a significant improvement were the following,

Was the marker's feedback returned promptly enough?
How did you feel about the amount of negative feedback?
How was the feedback in terms of detail?
Does the feedback you received mean you will change the way you produce your reports in future?
Overall did you find the feedback helpful?

The questions that showed no significant difference between the questionnaires were,

Did you receive the right amount of feedback?
How did you feel about the amount of positive feedback?

It is not surprising more students felt feedback was returned promptly enough with the second report as it was returned faster than the first. With regards to amount, even though there was no significant improvement the number of students satisfied with this before the intervention was relatively high (73.5%). There being no significant difference might only reflect the small non significant improvement in the amount of positive feedback. Giving useful positive feedback is one of the most difficult parts of effective feedback. This is especially true for Mathematics related subjects where it is often the norm for positive feedback to consist entirely of ticks. On the other hand the amount of negative feedback had improved.

The most promising improvement is the increase in detail in the feedback. Butler (1987) observed that precise feedback keeps attention focused on the task rather than the self. When attention is directed to the task performance is enhanced. The questionnaire also gave students the opportunity to explain what they would change, if anything, because of the feedback. On both reports there were a large number of students who needed to justify their answers more and be more precise with the Statistics phraseology they were using.

An updated version of the questionnaire was implemented in the Statistics class for Psychologists and Social Scientists, following the return of their first mini project. 90.9% of the students were satisfied with how promptly the project was returned. Compared with the second year students this figure is in between the figures for before (81.6%) and after (96.8%) the intervention.

The students were most unsatisfied with the detail of the feedback. This is most likely because the postgraduate markers are given no guidance on how to produce feedback. The most common reason given for why the feedback was not detailed enough is “There was no suggestion for improvement.” One of the key guiding principles is to keep criticism without suggestion for improvement to a minimum. This helps negative feedback to be seen as more than a criticism as well as ensuring students can progress.

The next most frequent reason given was “It was unclear where the mark was lost.” The reports unclaimed by students that were examined made students complaint here immediately clear. Some markers’ only written comment was the mark and if this was not full marks it was impossible for the student to know where they went wrong. Even though the markers received no instructions it would appear obvious this is bad feedback. Wilson (1999) found detailed feedback reduced students concerns over the fairness of their mark.

A test of association revealed that how satisfied students are with detail and amount of feedback are connected. This means if there was not enough feedback there was generally also not enough detail and vice versa. Tests of homogeneity were carried out to find out if there were any differences, for three of the questions, between the lab markers. This showed the responses were not significantly different for the amount, detail or usefulness of the feedback. However informally, there appeared to be some differences but every marker is bound to differ slightly.

These findings resulted in the production of a practical set of guidelines for markers. This was a reduced version of the guiding principles from the literature. The guidelines that were relevant and useable by markers were kept and real examples of good and bad feedback were inserted. The examples should help make the guidelines clearer and can be applied when writing feedback. Analysing the feedback on past reports and projects revealed a significant

problem. How much thought and time each tutor puts into marking is very different. This means some students will be receiving much more helpful feedback than others. It may be the department will have to provide equal training on marking work as at the moment the tutors receive nothing.
Even though the guidelines included in the practical advice to markers were examples from non-Statistics literature, it appears they can successfully be applied to assessment tasks in Statistics. This is because of the varied assessment tasks Statistics now uses. These tasks are now much more similar to those involved in non-Mathematical subjects. Generalising already successful feedback guidelines to Statistics is in itself a major outcome of this project.

The briefing document will still require a proper trial with tutors. Similar to the pilot in this study students could give their opinions before and after their tutors were given the guidelines. Another possibility would be a parallel groups design. This would involve some tutors being trained using the briefing document while others are not. This could have been incorporated into the present study but unfortunately this study's main limitation was the small number of participants. In any future work it would be really useful to increase the sample sizes. Ideally this would involve more than one institution. This would increase the overall sample size as well as allowing the guidelines to be tested in different contexts.

6.2 Self esteem, Positive Labels and mindsets

The updated questionnaire used in our work included a question on how much attention students think they pay to feedback in relation to others and a measure of students self esteem. From the results it appeared these concepts could be related. The students who claimed to pay least attention to feedback had lower self esteem scores, on average, than other groups. Young (2000) found that the students who see feedback as a criticism are those with low self esteem. This fits with this study’s results. No one wants to be criticised so those with low self esteem may tend to avoid feedback if that is how they perceive it.

Multiple comparisons applied to our data revealed that the students who paid the same and more attention than the average student, both have a mean self esteem score that is significantly higher than those of students who pay less attention than the average student.

Self esteem was the only personal characteristic of students that we attempted to measure and assess in this study. After the fieldwork with students had been completed, our attention was drawn to the broader-based work of Carol Dweck (Dweck, 2006), in particular her concerns over potential dangers of positive labels and the effects of student mindsets.

In our work with Statistics 1C, 26% more students wanted further positive feedback compared to negative feedback. This contrasts with the finding that students often think positive feedback is meaningless as it gives them nothing to develop on. However positive feedback being useless may be the least of a student’s worries. A study using a non verbal IQ test found some positive labels can be dangerous to the student (Dweck, 1999). The students were given ten problems and some of them were praised at the end. Some of the praise was on ability. For example “Wow you got eight right that’s a really good score you must be smart at this.” Other pieces of praise concentrated on effort. For example “Wow you got eight right that’s a really good score you must have tried really hard.” After this the students were offered a more challenging test they could learn from. Students that received the ability praise rejected the opportunity. However 90% of those who heard the effort praise wanted the challenge. All students were then given harder problems where they did not perform as well. The ability group interpreted this as they were not smart. This is not surprising as they were told doing well on the tests equalled being smart. The effort group did not take their worse performance as reflecting intelligence. They took this to mean they should try harder. The worrying part was that the ability group's performance dropped even when they returned to the easy problems. With the effort group their performance increased. The unfortunate conclusion here is praise emphasising ability decreases students IQ scores. 40% of the ability group also lied about their scores when they were asked at the end of the tests. Not only did these ability positive labels damage performance but they made a significant amount of students feel they had to lie to cover up a worse performance. The lesson here is that positive labels can be very dangerous when they suggest ability is a latent trait that can not be changed. Further study will also be required into avoiding the damaging effects of wrongly-labelled praise when constructing feedback. This will be a really useful addition to the guiding principles for constructing feedback.

Likewise negative feedback can be both useful and damaging. In another study by Dweck (2006), students were given critical but helpful feedback. Most of Carol Dweck's work concentrates on her theory of mindsets. Dweck proposes everyone has one of two mindsets. These are the growth and fixed mindsets. A growth mindset means the person believes that with motivation and effort you can become better at anything. A fixed mindset means the person believes that ability is a fixed trait. In her study on negative feedback, before the feedback was given, the evaluator was made out to be arrogant. The majority of the fixed mindset students took the constructive criticism as an insult and blamed the evaluator. On the other hand the students with a growth mindset concentrated on the feedback and felt it was honest and that they will learn from the evaluator. This may explain the conflicting results that negative feedback reduces motivation and students want and need this type of feedback to improve. Perhaps the results do not conflict at all but it depends on the students' mindsets.

One of Dweck's studies (2002) involved giving students feedback while measuring their brainwaves. Those with fixed mindsets stopped paying attention after they found out if their answer was correct or incorrect. The brain waves showed no attention was paid to the feedback that would improve their performance. They were not even interested in finding out the correct answer for the questions they answered wrong. Those with the growth mindset showed a different response. These students paid attention to all feedback.

Another study by Dweck (2006) used seventh graders. Some of these students had received a poor grade in a new course they were taking. This motivated those with a growth mindset to study harder for the next test. On the other hand the fixed mindset students studied less. Their perspective was that, if you do not have the capability, then why misuse your time on it. Instead of learning from their failures they appeared to only care about mending their self esteem. For example college students were given the opportunity to look at other students answers after a failure. Those with a growth mindset looked at students who had done better than them as they want to improve. However the fixed mindset students picked students who had performed worse then them. This would make them feel better and repair their self esteem.

In Dweck's studies it is the fixed mindset students who pay less attention and concentrate on repairing self esteem. In this study the students who pay less attention appear to have a corresponding low self esteem. It is likely these students have a fixed mindset and repairing self esteem may be so important to them since they have little of it to begin with. However if there is a relationship between attention paid to feedback and self esteem it is not clear what direction this is in. It could be the students with low self esteem do not pay that much attention to feedback. On the other hand it may be that not paying enough attention to feedback actually lowers your self esteem. One possible explanation is by not reading any feedback in detail the comments may initially seem personal. Future work could measure students' mindsets as well as their self esteem and attention paid to feedback. This may show evidence of a more complex relationship than we found in this study.

Another issue with our survey was the way attention paid was measured. This study's results only revealed how much attention the students think they pay and not how much attention they actually pay to feedback. There have been studies that have measured the latter. For example Higgins et al (2002) found 39% of students spend less than five minutes reading feedback and 81% spend less than fifteen minutes. This means the average student in Higgins study spends less than fifteen minutes on feedback. It is difficult to compare this study and Higgins since in this study the students were not asked about time spent. They were asked to compare themselves to the average student but every student could have had a different concept of the average student. Future work could ask students specifically about the time they spend or it would be more accurate to actually time students going over their feedback. Asking students how much attention they pay to their feedback is obviously subject to bias. Some students may want to portray themselves as a good student who pays extensive attention to feedback. On the other hand some may state they pay little attention to highlight how useless they think the feedback is. Observing the students with their feedback would avoid this problem. However this method would not be completely accurate either. Students would be observed in the classroom but they may return to their feedback at a later time. Perhaps if the self report measure was more specific the bias could be reduced. Students could be asked roughly how long they spend on their feedback. This way students are unaware of how long the average student spends on their feedback.

Self esteem also appeared to differ between the students who felt their mark was fair and those who felt it was unfairly poor. This result makes perfect sense since having a high opinion of yourself often means you think your work is better than it really is.

6.3 Instant feedback with multiple choice testing

Model Choice, the department's computer based multiple choice assessment system that tests knowledge of probability models, received a small trial with four students. The results of this were very positive. All of the students strongly agreed that the system was easy to use, that they appreciated the immediate feedback and that the feedback means they will be able to solve similar problems. The only question that received any disagreement was 'The correct answer was obvious after the initial feedback?' The students' answers varied between strongly agree, slightly agree and slightly disagree. This was a good result as we did not want all of the students agreeing or disagreeing with the statement. If all the students agreed the answer was made obvious then the second attempt would be essentially useless. On the other hand if all the students disagreed then that would suggest the feedback was unhelpful. The responses to the open ended questions were also largely positive. The students found it difficult to think of any negative comments about the system. However this may be because this testing was on a one to one basis and the students may have felt uncomfortable criticising the system. Ideally Model Choice would have been trialled with more participants.

When constructing the multiple choice questions for the S1C quiz we tried to follow the appropriate guidelines as closely as possible. The guidelines that are clearly visible in the program are each question refers to a specific principle in the course, use a mixture of type of question, base incorrect answers on common mistakes and avoid 'all of the above' and 'none of the above'. We also managed to follow the guidelines that the options should be random and options should be presented in a suitable order, which seems contradictory, to an extent. These guidelines appeared in different articles but both provide useful advice. Obviously it is impossible to follow both guidelines on the same question. However we found that depending on the question one of these guidelines would be more suitable. For the questions with increasing length options it made sense for these to be ordered shortest to longest. For these questions a random arrangement would only make the options more difficult to read. On the other hand for the rest of the questions the options being random means no pattern, even an unintended one, can be seen through the questions. For example some people constructing multiple choice tests avoid putting the correct answer at position A. They feel that it is too easy and students may not even read the rest of the options. Randomness avoids this being a problem. One guideline we did not follow was keeping questions independent from each other. This is because linking some questions can be an advantage to the students. When two questions use the same context reading is reduced for the student and more time can be spent on the Statistics. The benefit of keeping questions independent is not clear.

Unfortunately there are no guidelines on the most important feature of the quiz, the feedback. Without feedback on why answers are incorrect or correct there is little benefit of the quiz as formative assessment. Therefore we created our own immediate feedback. We found that the crucial guideline should be achieving a balance between giving helpful feedback but not giving the answer away. This was quite challenging and the biggest problem we came up against was creating the feedback for the ‘do not know option’. This was because when students choose ‘do not know’ it is impossible to know which part of the question they do not understand. We decided the most useful feedback here would be a broad hint. Sometimes this was the same as the feedback for one of the incorrect answers. In the future these ideas could be incorporated into a set of guiding principles for constructing feedback for multiple choice tests.

6.5 Comparing Staff and Students Responses

Next two of the labs were piloted with some of the staff and postgraduates in the department, then with some third year Statistics students. Both of these groups contained twelve participants. Participants were instructed to work through the appropriate lab and then fill out the adapted Model Choice Questionnaire. All of the participants in both groups strongly agreed that the system was easy to use. With the second statement, ‘the program is a good way of assessing learning’, the staff’s responses were more positive than the students. The majority of the staff strongly agreed with the statement where as the students responses varied between strongly agree, slightly agree and do not agree or disagree. This could be because from a students perspective the easier the summative assessment the better. From observing the students some of them seemed to find the questions quite difficult. However the quiz will be given to students immediately after the lab so it was essential that the questions were not too easy. The staff also showed more agreement with ‘there was enough detail in the feedback’ than the students. Again the students may be focusing on what can improve their mark. Both groups mainly agreed that the feedback was easy to understand. The next statement, ‘the correct answer was obvious after the initial feedback’, could have been difficult to answer. The rest of the statements are all positive aspects of the system but it is not so clear cut with this statement. Participants may either think the purpose of the feedback is to make the answer clear or the answer is not supposed to be obvious. This will obviously affect participants’ responses. For both groups a significant amount of participants did not agree or disagree with the statement. This is a good result since we wanted to avoid too much agreement or disagreement. The last statement, ‘getting feedback means students will be better able to solve similar problems’ also received similar responses from both groups, with all participants agreeing to some extent. The open ended questions showed a lot of support for the plausibility of the incorrect options and the usefulness of the feedback. Some of these responses actually led to real changes being made to the system. For example, a couple of members of staff pointed out that how the correct answer is reached is not explained, only given, if a student answers incorrectly on both of their attempts. We were too busy concentrating on explaining the correct answer to those who choose it we overlooked the explanation’s absence elsewhere. Explaining how the correct answer is reached is now always present beside the correct answer.

The next criticism which lead to an improvement to the system came from the students. Many of the students complained that the question and options were not spaced out enough. We discussed this with the students and decided to reduce the font and increase the spacing. Finally due to the students' suggestions we created a summary page which showed their results for each question type. The advantage of this is students can immediately see their strengths and weaknesses. Having two trials, with staff and students, was a very worthwhile exercise as both groups had different perspectives on the system.

The main achievements of this project were the development of specific guidelines for feedback in Statistics assessments and an online assessment with feedback that can be used for both formative and summative assessment. The next stage will be introducing the quiz into the relevant S1C labs. Unfortunately this could not be part of the present project as the funding was only for one year.

Appendix A

Feedback Questionnaire

1. What did you receive on this assignment out of 20?

0-5 □ 6-10 □ 11-15 □ 16-20 □

2. How did this result fit with your expectations?

Worse then expected □ As expected □ Better than expected □

3. Was the feedback returned promptly enough?

Yes □ No □

4. Did you receive the right amount of feedback?

Yes □ No, too little □ No, too much □

5. How do you feel about the amount of negative and positive feedback? (Tick more than one if applicable)

Not enough negative □ Not enough positive □

Right amount of negative □ Right amount of positive □

Too much negative □ Too much positive □

6. How was the feedback in terms of detail?

Not detailed enough □ Right amount of detail □ Too detailed □

Give an example of this from the feedback you received in the box below.

7. Did the feedback you received mean you will change the way you produce your reports?

Yes □ No □

Explain why and how in the box below.

8. Overall, did you find the feedback helpful?

Yes □ No □

Thank you for taking the time to complete this questionnaire.

Feedback Questionnaire Follow Up

1. Which (if any) of our Level 1 Statistics courses did you take in a previous year?

Stats 1Y/1Z □ Stats 1B □ Stats 1C □ None □

2. What mark did you receive on this assignment (out of 20)?

0-5 □ 6-10 □ 11-15 □ 16-20 □

3. How did this mark compare with your mark for the last lab report in the same course?

Worse than before □ About the same □ Better than before □

4. Was the marker's feedback returned promptly enough?

Yes □ No □

5. Did you receive the right amount of feedback?

Yes □ No, too little □ No, too much □

6. How did you feel about the amount of negative and positive feedback? (Tick more than one if applicable)

Not enough negative □ Not enough positive □

Right amount of negative □ Right amount of positive □

Too much negative □ Too much positive □

/OVER

7. How was the feedback in terms of detail?

Not detailed enough □ Right amount of detail □ Too detailed □

Please give an example of this from the feedback you received.

8. Did you make changes to the way you produced this report based on the marker's feedback last time?

Yes □ No □

9. Does the feedback you received this time mean you will change the way you produce your reports in future?

Yes □ No □

Please explain why or why not, and describe what you will change.

10. Overall, did you find the feedback helpful?

Yes □ No □

Thank you for taking the time to complete this questionnaire.

Appendix B

Briefing Document - Feedback Guidelines

This is a summary of the literature on what constitutes good feedback. Examples of feedback given to the Statistics class S1C have been included to illustrate the main points.

Balancing Positive and Negative

Negative feedback itself can have damaging consequences but it can also be a method for improving students' performances. An important negative effect of criticism is the reduction in confidence and motivation in students (Taylor and Hoedt, 1996). For example criticism can be seen as a confrontation which serves to challenge a student's confidence. Many teachers are aware of this and keep criticism without any suggestion for improvement to a minimum. Here are some examples of markers following this rule.

> "This gives a wrong representation of the populations responses. Percentages should be within levels of region" – Report N2, student received 6/10.

> "This is a wrong test. Remember you are comparing two populations you should have used 2 sample t test" – Project 1, student received 10/25.

However not all of the feedback is explicit about what should be changed.

> "You have treated age as a categorical variable here" – Report L3, student received 9/10.

In Hyland and Hyland's (2001) study 76% of negative feedback was in some way made less severe. They found many teachers used imprecise quantifiers such as 'some' and 'little' to mitigate criticism.

Despite its short comings negative feedback is a critical tool for helping students to realize and improve their weaknesses. The purpose of feedback is to provide an accurate account of how good the work is. This would be virtually impossible without any negative feedback. In a study at Nottingham Trent University (Weaver, 2006) a tremendous majority felt 'constructive criticism is needed to know how to improve.' This was 100% of the Art and Design students and 92% of the Business students. The crucial thing is that the criticism is constructive and delicately managed. It has been found that students are most motivated when their goals are not too difficult to achieve (Freeman and Lewis, 1998). This suggests all constructive criticism should be seen as attainable by the student.

With regards to positive feedback, intuitively praise for good work will result in an increased positive attitude (Gee, 1972). Many students can become pessimistic about their work and disregard the feedback if no positive comments are given. Giving feedback on what a student has done well is essential for knowing what to repeat in future work. Weavers study (2006) showed strong agreement from students that more praise should be provided.

An ideal structure is the sandwich, where criticism is sandwiched between two pieces of positive feedback. It is also strongly recommended that the feedback avoids the word ‘but’ which can devalue the praise being given (Brockbank and McGill, 1998).

> “Good but your descriptive analysis was incomplete” – Project N2, student received 6/10.

If positive statements are seen as insincere it is doubtful they will be motivating. The next example illustrates a useful method of giving a summary comment.

> “Very good. I hope the comments above will be helpful for future reports” – Report M3, student received 8/10.

It is effective because it ends with positive comments but a reminder to study the previous constructive criticisms. Brophy (1981) found that for positive feedback to be effective it needs to be informative and realistic.

Detail

General statements are of no use and all feedback should be specific (Brockbank and McGill, 1998). No benefit is taken from comments such as, “good piece of work” or “not good enough”. A study which interviewed students at Robbins University found widespread disappointment with how detailed the feedback was. Much of this focused on how little detail they were given on how they could improve their work (James, 1996). The subsequent example illustrates this.

> “2 groups, continuous, in range of etc” – Report F3, student received 7/10.

It seems unlikely the student would know exactly what to change from this statement. Another problem arises with the use of questions.

"Randomness of samples?" – Report 4, student received 8.5/10.
"Effect of age?" – Report L2, student received 9/10.

Many students are confused by these questions and do not identify them as negative feedback which requires them to change something. Furthermore the examples do not contain enough information on where marks were lost. These questions would be much more helpful phrased as statements such as the following.

"You must state your reasons for saying this" – Project L2, student received 8/10.

This statement is much more helpful than the frequently written "Why?" The situation appears even worse when the work is good. General comments are more commonly found with praise and students find this very frustrating (Cowan, 2006).

"Very good" - Report L2, student received 9/10.

The following is a much better example as the marker specifies exactly where the praise is aimed at and explains why this was good work.

This is an excellent plot for question 2 it shows that the two data sets are dependent" – Project 1, student received 10/25.

When feedback is made specific there is even a reduction in students concerns over the fairness of their mark (Wilson, 1999). It seems that once provided with the correct information students can understand their mistakes and agree with the marker. Several researchers go as far as to say that specificity is correlated with learning.

Amount

Many believe that feedback should only focus on a few areas so that students know exactly what to change. This is especially true with negative feedback and for maximum benefit it may have to be limited to one or two areas (Brockbank and McGill, 1998). After this many students switch off so it is better not to concentrate on insignificant or infrequent errors (Wilson, 1999). However Weaver (2006) discovered that 96% of Business students and 75% of Art and Design students felt they were not provided with enough feedback.

Objective

One way markers can be objective is by phrasing their comments as their own view.

“I think the aim has to do with testing assumptions of normality” – **Body Fat,** student received 7.5/10.

This shows their feedback may not be a universally agreed view and takes accountability for what they are saying (Brockbank and McGill, 1998). This is also a useful method for softening criticism. It acts to decrease the authorative gap between teacher and student. This may mean that students are less threatened and more willing to take the feedback on board (Hyland and Hyland, 2001).

Timely

The most repeated principle for effective feedback in the literature is that it must be timely. Feedback can never come too soon and should be given as soon as possible. If the delay in receiving feedback means the student is on a different part of the course, it is unlikely any real attention will be paid to the feedback. Hartley and Chesworth in 2000 found 59% of students felt feedback was given too late to be helpful. Even if attention is paid, very few students will enquire about their feedback if the delay has been long (Filer, 2000). The principle of timely feedback is thought to be even more important for first year students. For these students it is essential that before more work is completed, they know what to change or are given the confidence they deserve (James, 1996).

Future Oriented

Feedback should have clear implications for the current and future tasks. Therefore feedback should not be limited because the work is a final draft. Even though nothing can be changed on the current work the advice can be taken to forthcoming assignments. The marker should be doing more than justifying the grade for the current project. If feedback is written with this principle in mind recurring problems should be evident to the student. The following are perfect examples of this, as the markers are giving constructive feedback even though the student received full marks.

“Next time you can reduce the font size so that everything will be accommodated” – Report F3, student received 10/10.

“”Safer to use the probability plot to establish normality” – Report F3, student received 10/10.

References

Ammons R.B. (1956) Effects of knowledge of performance: A survey and tentative theoretical formulation. *Journal of General Psychology* **54,** 279-299.

Audia, P.G., Locke, E.A., Smith, K.G. (2000) The paradox of success: An archival and a laboratory study of strategic persistence following a radical environmental change. *Academy of Management Journal,* **43,** 837-853.

Baldwin B.A. (1984) The role of difficulty and discrimination in constructing multiple choice examinations: with guidelines for practical application. *Journal of Accounting Education* **2,** 19-28.

Bandura A., & Cervone D. (1983) Self-evaluative and self efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology* **45,** 1017-1028.

Baron, R.A. (1993) Criticism (informal negative feedback) as a source of perceived unfairness in organizations: Effects, mechanisms and countermeasures. In Greenberg (Ed), *Justice in the workplace: Approaching fairness in human resource management.*

Begley C.M., & White P. (2003) Irish nursing students' changing self-esteem and fear of negative evaluation during their preregistration programme. *Journal of Advanced Nursing* **42,** 390-401.

Black, P., & William, D. (1998) Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice,* **5,** 7-74.

Blayney, P., & Freeman, P. (2003) Automated marking of individualised spreadsheet assignments: The impact of different formative self-assessment options. *In Proceedings of the 7th international computer assisted assessment conference,* Loughborough.

Brockbank, A. & McGill, I. (1998) *Facilitating reflective learning in higher education* (Buckingham, The Society for Research into Higher Education & Open University).

Brophy, J. (1981) Teacher Praise: A functional analysis. *Review of Educational Research,* **51,** 5-32.

Buchanan, T. (2000) The efficacy of a World-Wide Web mediated formative assessment. *Journal of Computer Assisted Learning* **16,** 193-200.

Butler R. (1987) Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology* **58,** 1-14.

Carless, D. & Keppell, M. (2006) Peer feedback: the learning element of peer assessment. *Teaching in Higher Education,* **11,** 279-290.

Cizek G. J., & Rachor R.E. (1995) *Nonfunctioning options: A closer look.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Charman, D., & Elmes, A. (1998) A computer-based formative assessment strategy for a basic statistics module in geography. *Journal of Geography in Higher Education,* **22,** 381-385

Cowen, J. (2006) *On Becoming an Innovative University Teacher* (Berkshire, The Society for Research into Higher Education & Open University).

Crisp B.R. (2007) Is it worth the effort? How feedback influences students' subsequent submission of assessable work. *Assessment & Education in Higher Education* **32,** 571-581.

Crisp, V., & Ward, C. (2007) The development of a formative scenario-based computer assisted assessment tool in psychology for teachers: The PePCAA project. *Computers and Education,* **50,** 1509-1526.

Dempsey, J.V., Driscoll, M.P., & Swindell, L.K. (1993) Text-based feedback. In J.V. Dempsey, & G.C. Sales, *Interactive instruction and feedback* (pp, 21-54). NJ: Educational Technology Publications.

Dragga, S. (1986) *Praiseworthy grading. A teacher's alternative to editing error.* Paper presented at the conference on College Composition and Communication, New Orleans.

Downing S.M., Dawson-Saunders B., Case S.M., & Powell R.D. (1991) *The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II item characteristics.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Dweck, C. S. (1999a). Caution-praise can be dangerous. *American Educator, 23(1),* 4-9.

Dweck, C.S. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In J. Aronson (Ed.), *Improving academic achievement.* New York : Academic Press.

Dweck, C. (2006) *Mindset: The New Psychology of Success***.** (Random House).

Earley P.C. (1988) Computer generated performance feedback in the magazine subscription industry. *Organizational Behavior and Human Decision Processes* **41,** 50-64.

Freeman R., & Lewis R. (1998) *Planning and Implementing Assessment* (London, Kogan Page Limited).

Fritz C., Morris P., & Bjork, R. (2000) When further learning fails: stability and change following repeated presentation of text, *British Journal of Psychology,* **91**, 493–511.

Fry H., Ketteridge S., & Marshall S. (2003) *A Handbook for Teaching and Learning in Higher Education* (Oxon, RoutledgeFalmer).

Gal I., & Garfield J.B. (1997) *The Assessment Challenge in Statistics Education* (Ohmsha, IOS Press)

Gammon, J., & Morgan-Samuel, H. (2004) A study to ascertain the effect of structured student tutorial support on student stress, self-esteem and coping. *Nurse Education in Practice,* **5,** 161-171.

Gee, T.C. (1972) Students' responses to teacher comments. *Research in the Teaching of English,* **6,** 212-221.

Gipps, C.V. (2005) What is the role of ICT-based assessment in universities? *Studies in Higher Education,* **30,** 171-180.

Haladyna T.M., & Downing S.M. (1993) How many options is enough for a multiple choice item? *Educational and Psychological Measurement,* **53,** 999-1010.

Haladyna T.M., Downing S.M., & Rodriguez M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education,* **15,** 309-334.

Hartley, J. & Chesworth, K. (2000) Qualitative and quantitative methods in research on essay writing: no one way, *Journal of Further and Higher Education,* **24**, 15–24.

Higgins, R., Hartley, P., & Skelton, A. (2002) The Conscientious Consumer: reconsidering the role of assessment feedback in student learning. *Studies in Higher Education,* **27,** 53-64.

Hirsch M.L., & Gabriel S.L. (1995) Feedback Strategies: Critique and evaluation of oral and written assignments. *Journal of Accounting Education* **13,** 259-279.

Howe K.R., & Baldwin B.A. (1983) The Effects of Evaluative Sequencing on Performance, Behaviour and Attitudes. *The Accounting Review* 135-142.

Hounsell, D. (1987) Essay writing and the quality of feedback in: JTE Richardson, MW Eysenck & D. Warren-Piper (Eds) *Student learning:*

research in education and cognitive psychology. (SRHE & Open University Press)

Huntley, R.M., & Welch, C.J. (1993, April) *Numerical answer options: Logical or random order?* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Hyland F., & Hyland K. (2001) Sugaring the pill: Praise and criticism in written feedback. *Journal of Second Language Writing* **10,** 185-212.

Ilgen D.R., Fisher C.D., & Taylor M.S. (1979) Consequences of individual feedback on behaviour in organization. *Journal of Applied Psychology,* **64**, 349-371.

James, D. (1996) 'Mature Studentship in Higher Education', PhD thesis, University of the West of England.

James, D. (2000) *Assessment: Social Practice and Social Product*

Kluger A.N., & DeNisi A. (1996) The effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin* **199,** 254-284.

Kluger A.N., & DeNisi A. (1998) Feedback Interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science* **7,** 67-72.

Komaki J., Henizmann A.T., & Lawson L. (1980) Effect of training and feedback: Component analysis of a behavioral safety program. *Journal of Applied Psychology* **65,** 261-270.

MacLellan E. (2001) Assessment for learning: the differing perceptions of tutors and students, *Assessment and Evaluation in Higher Education,* **26**, 307–318.

McMorris, R.F., Boothroyd, R.A., & Pietangelo, D.J. (1997) Humor in educational testing: A review and discussion. *Applied Measurement in Education,* **10,** 269-297.

Mikulincer M. (1990) Joint influence of prior beliefs and current situational information on stable and unstable attributions. *Journal of Social Psychology* **130,** 739-753.

Moreno, R., Martinez, R.J., & Muniz, J. (2004) Guidelines for the construction of multiple choice items, *Psicothema,* **16,** 490-497.

Nunnally J.C. (1970) *Introduction to Psychological Measurement*, (McGraw-Hill, 1970) 209-212.

Orsmond P., Merry S., & Reiling K. (2005) Biology students' utilization of tutors' formative feedback: a qualitative interview study. *Assessment & Education in Higher Education* **30,** 369-386.

Pressy S.L. (1950) Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self instruction. *Journal of Psychology* **29,** 417-477.

Pritchard R.D., Jones S.D., Roth P.L., Stuebing K.K., & Ekeberg S.E. (1988) Effects of group feedback, goal setting and incentives on organizational productivity. *Journal of Applied Psychology* **73,** 337-358.

Pulford B D., Johnson A., & Awaida M. (2005) A cross cultural study of predictors of self-handicapping in university students. *Personality and Individual Differences* **39,** 727-737.

The Quality Assurance Agency for Higher Education (2006): Code of practice for the assurance of academic quality and standards in higher education, Section 6: Assessment of students, Second Edition, The Quality Assurance Agency for Higher Education.

Ramaprasad A. (1983) On the definition of feedback, *Behavioural Sciences,* **28**, 4–13.

Roberts D.M. (1993) An empirical study on the nature of trick test questions. *Journal of Educational Measurement,* **30,** 331-344.

Rogers W.T., & Harley D. (1999) An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement,* **59,** 234-247.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rust C. (2002) The impact of assessment on student learning, *Active Learning in Higher Education,* **3**, 145–158.

Smith G. (2007) How does student performance on formative assessment relate to learning assessed by exams. *Journal of College Science Teaching* 28.

Smith, E., & Gorard, S. (2005) 'They don't give us our marks': the role of formative feedback in student progress. *Assessment in Education: Principles, Policy and Practice,* **12,** 21-38.

Taylor, W.F., & Hoedt, K.C. (1966) The effect of praise upon quantity and quality of creative writing. *Journal of Educational Research,* **60,** 80-83.

The National Student Survey (2006) Commissioned by the Higher Education Funding Council for England.

Thelwall, M. (2000) Computer-based assessment: a versatile educational tool. *Computers & Education,* **34,** 37-49.

Vallacher, R.R., & Wegner, D.M. (1987) What do people think they're doing? Action identification and human behaviour. *Psychological Review,* **94,** 3-15.

Vollmeyer R., & Rheinberg F. (2005) A surprising effect of feedback on learning. *Learning and Instruction* **15,** 589-602.

Weaver M.R. (2006) Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Education in Higher Education* **31,** 379-394.

Wilson D. M. (1999) Improving feedback on student papers: a quantitative method which aids marking and gives valid feedback. *New Education Today* **11,** 53-56.

Wise S.L., Plake B.S., Eastman L.A., Boettcher L.L., & Lukin M.E. (1986) The effects of item feedback and examine control on test performance and anxiety in a computer-administered test. *Computers in Human Behavior* **2,** 21-29.

Young P. (2000) ‘I might as well give up’: self-esteem and mature students’ feelings about feedback on assignments, *Journal of Further and Higher Education,* **24**, 409–418.