Assessment and Feedback

2019 version by **Jon Guest**,
Aston University
Edited by **Caroline Elliott**
and **Matthew Olczak**,
Aston University
Published August 2019

Contents

1. Introduction

1.1 Background

It is difficult to overestimate the impact assessment has on learning. For the majority of students it determines (a) the topics on the syllabus they choose to study (b) how much time they spend studying them and (c) how they go about studying them. For example, do they simply try to memorise the course content or do they attempt to understand the material? Unfortunately, tutors tend to spend most of their time thinking about module content and delivery i.e. how to run large and small group teaching. Assessment considerations tend to be something of an afterthought. One possible cause of this bias is that academics are not representative of their classes: they are much more intrinsically motivated in the subject than the majority of people they teach.

Given its importance, it is worrying that evidence from the National Student Survey suggests that students are less satisfied with assessment and feedback than any other aspect of their educational experience in higher education. For example, in 2018, 86 per cent of economics students agreed that 'staff are good at explaining things'. This figure falls to 66 per cent for those who agreed that 'I have received helpful comments on my work'.

One reason for this poor experience is clearly increasing student numbers. Many institutions have been able to exploit economies of scale in delivery by simply increasing class sizes i.e. by spreading fixed and 'lumpy' costs. Evidence from the NSS suggests that most universities have done this whilst maintaining student satisfaction with the quality of teaching. Assessment and feedback, however, has a higher proportion of variable costs. The resources required to grade and provide traditional written feedback tend to increase proportionally with the number of students on the module.

There is also less innovation in assessment than teaching. This is probably because of the greater perceived risk. Tutors are more likely to trial different methods of delivery as, if they are unproductive, it only affects one or a small number of classes. However, if new and different types of assessment prove ineffective or detrimental to student performance, the potential impact is much greater and longer lasting.

1.2 The objectives and purpose of assessment

Assessment is complex and has a number of competing objectives. These include:

a. *Promoting/supporting learning*: Assessment design should support and encourage effective and deep learning. This is often referred to as 'assessment **for** learning'

b. *Measurement*: Another purpose of assessment is to measure (a) what students have learnt on the module/course and (b) the level/depth of this understanding. It judges the extent to which students have achieved the learning outcomes for the module and programme. This is often referred to as 'assessment **of** learning'.
c. *Providing feedback for tutors*: Assessment performance helps tutors to evaluate the effectiveness of their own teaching. A relatively low distribution of final grades and/or poor feedback comments on student questionnaires should trigger some reflection by the module leader.
d. *Identifying students*: Assessment performance can help to identify those students who are struggling to meet learning outcomes and may need extra support and guidance.

Module leaders need to keep these objectives in mind when thinking about assessment. They also need to take into account the manageability of the process. The two most important costs for a tutor are the time it takes to (a) write assessment questions, and (b) grade and provide feedback.

Section 2 of this handbook chapter focuses on assessment for learning in more detail while section 3 examines some implications of assessment design within a module i.e. the number/ structure of coursework and type of examination. Section 4 considers the implications of using different types of assessment question and, given their popularity, discusses the use of multiple-choice questions in some detail. It also includes cases studies that outline some alternative types of assessment. Section 5 concentrates on methods of feedback – the area of assessment where students express the least satisfaction with their experience in higher education.

Section 6 of the handbook focuses on two important issues when considering the measurement function of assessment – validity and reliability. The validity of an assessment is the extent to which it measures what it purports to measure, i.e. students' understanding of module content. For example, the extent to which guesswork as opposed to knowledge can influence grades. The reliability of an assessment is the extent to which grading is consistent both between different assessors (inter-marker reliability) and by the same assessor (intra-marker reliability). The chapter provides some tips on how to improve both reliability and validity, but recognises the trade-offs that exist.

2. Assessment for learning

One of the key functions of assessment is to facilitate, motivate and support learning – assessment for learning. Assessment can support learning in two important ways.

Firstly, there is the process of completing the work i.e. the research, reading, writing and revision. Therefore, both the design within a module (the number of coursework vs exam) and the type of question (fixed vs open response) play a key role in promoting and determining the level/depth of learning.

Secondly, there is the communication between marker and student about the quality of work both during and after its completion. This includes (a) information on strengths and weaknesses and (b) advice on how to improve performance in the future. For feedback to be effective, students must act upon it.

Section three of this chapter discusses some of the implications of assessment design within a module while section four discusses different styles of question and types of assessment. Section five discusses feedback in more detail.

2.1 The importance of assessment

It is difficult to overestimate how important coursework, tests and exams are in the learning process. Some widely cited studies from the 1970s concluded that assessment was by far the most important factor that determined students' study time. Snyder (1971) found that students differentiated between the actual curriculum and the hidden curriculum i.e. what they needed to know to perform well in graded work. Miller and Parlett (1974) coined the term 'cue seekers' to describe those students who go to great lengths to find out the best way to answer a question or what topics they need to learn for the examination. More recently, Thomas, Hockings, Ottaway and Jones (2015) found that the number of non-contact study hours depends on the perceived assessment demand of the modules. Chevalier, Dolton and Lurrman (2017) carried out an interesting study on a large first year Principles of Economics module that uses on-line quizzes to promote continuous learning. The authors found that making the quizzes count towards the final grade increases the participation rate by 42 – 62 percentage points.

Evidence and experience indicates that the majority of students concentrate on tasks where they see the most direct and obvious impact of their actions on their marks/grades i.e. a clear line of sight. The tendency to ignore non-graded tasks may also increase with experience e.g. final year undergraduates focus more on graded work than students in the first year of the course.

Therefore, assessment is very important as it potentially has a bigger impact on learning than the actual teaching on a module. Gibbs and Simpson (2004) argue that assessments should encourage:

- *An appropriate amount of study time*. The perceived demands of the assessment should incentivise students to exert enough effort so they can develop a deep understanding of the material.
- *A relatively even distribution of study time throughout the duration of the module*. Students will develop a deeper understanding of the material if they work consistently as opposed to a few hours or days of intensive study just before a coursework deadline or examination. Unfortunately, a combination of much larger student numbers and resource constraints make it very difficult for tutors to mark regular problem sheets, essays or other types of homework.
- *Study time on high-quality learning activities*. It is important that students do not perceive memorising and rote learning as effective ways to achieve high grades.

3. Assessment design to promote learning

How can different assessment design within a module help to address the conditions identified by Gibbs and Simpson (2004)? What are the advantages and limitations of the traditional closed book examination and what alternatives are available? How many pieces of coursework in a module are both desirable and feasible? This section of the handbook discusses these issues in more detail.

3.1 Traditional closed-book examinations

Traditional closed book examinations are by far the most widely used and heavily weighted assessment component on the majority of economics modules. It is difficult to set a time-constrained exam where students have to answer questions that test their understanding of the entire module content. Typically, questions only cover a subset of the curriculum so both the level and specificity of any guidance is an important factor.

If the lecturer provides very little advice about the topics the exam paper will cover, then ignoring any of the course content is a risky strategy for students. The fear of not learning an examined topic could incentivise risk averse students to work harder and more consistently throughout the module.

However, students' perceptions of the demands of the exam might not align with that of the tutor. They may believe it is possible to predict exam questions and so spend large amounts of time acting strategically[1] rather than focussing their efforts on mastery of the module content. This approach may also reduce the validity of assessment, as grades partly depend on luck.

If tutors provide more guidance, students may learn some topics in greater depth and gain a deeper understanding. However, there is a danger it reduces the perceived demands of the assessment, leads to lower effort levels and results in more inconsistent patterns of studying with students completely ignoring some topics. Some educational researchers also argue that traditional exams promote low quality learning activities such as attempts to memorise material and are a poor predictor of long-term learning and understanding of course content.

---

[1] This includes continual visits by some students in staff office hours with questions designed to reveal information about what is on the exam rather than about the understanding of module content.

3.2 Some alternative types of examinations

There are a number of innovative alternatives to the standard closed book examination.

Open Book Examinations

Open book examinations vary by the amount of material the assessor permits students to take into the exam room. In some cases, there are no limits. The students are free to take whatever materials they wish into the room. In other open book exams, the assessor places certain restrictions. Some examples include specified papers, books and limits on the quantity of notes i.e. one A4 sheet of notes.

One advantage with using this type of format is that it gives students more time to demonstrate higher order thinking skills as opposed to simple recall of basic information. It may also effectively signal that deeper learning is required to achieve high grades.

One potential disadvantage is the possibility it leads to less revision and exam preparation as some students overestimate how much they can effectively utilise books and notes in the exam.

Open Examinations

There are two broad categories of open examination.

a. The assessor gives students a task to complete in a much shorter time-period than coursework i.e. overnight or within a couple of days.
b. The assessor gives the student the assessment topic/questions and/or specific material (i.e. a case study/journal articles) to read and research before the exam. After a period for research/revision, students have to answer questions under normal exam conditions. If tutors provide detailed guidance for a closed book examination, there is a danger it effectively becomes this type of open examination but with lower expectations.

Open examinations can be useful if it is difficult to assess the module's learning outcomes in time-constrained conditions. For example, the ability to synthesise information from a wide range of academic sources. It is very important to communicate the length of time you expect students to spend on the preparation activities.

Case study 3.2.1: An example of an open examination

The tutor gave the following assessment instructions to students on an MSc Microeconomics module.

Select any firm of your choice. Discuss the microeconomic factors that you think impact on the firm's decision-making. Highlight any challenges the firm currently faces and how microeconomics helps us understand these challenges and possible strategies the firm could adopt to overcome them.

You must include relevant diagrams and/or mathematics in your answer. You cannot bring any notes or texts to the exam. However, if you have collected firm and industry relevant data in the form of tables or plots, these can be brought to the exam, discussed and submitted along with your test answer.

This format aimed to achieve a number of objectives. First, students needed to demonstrate a good understanding of the technical content of the module. Second, they had to ask themselves, on a weekly basis, how the material they were learning could provide insights into real firms and consumers' decision-making. Finally, they had the challenge of drawing together information on factors such as market share, pricing strategies and profitability whilst using microeconomic theory to provide insights.

Students were encouraged to show the tutor some of the materials they were planning to use to check they were suitable/appropriate. They were also advised to select a firm not being analysed by fellow students. This meant they could work collaboratively with their peers while confident that their exam answers would be very different from one another.

A number of exam scripts contained very high quality work i.e. demonstrating technical knowledge while producing detailed background research into a selected firm.

3.3 Coursework

Modules typically have one or two pieces of graded coursework to keep marking loads at a manageable level. One limitation with this approach is that students only study content they perceive as relevant for the assessments. It can also result in very inconsistent patterns of learning as studying in non-contact time takes place in short intensive bursts i.e. in the weeks/day prior to the submission deadlines.

To incentivise students to work more consistently, some type of continuous assessment is usually required. How is this possible on modules with large student numbers without creating an unmanageable marking load? There are a number of possibilities.

The module requires the submission of numerous assessments that are marked/graded by software rather than the tutors.

This approach is increasingly popular, with lecturers creating a series of multiple-choice tests/quizzes using on-line products such as MyLab Economics or Aplia. The students typically complete the tests outside the classroom and the software automatically grades the assessment and provides feedback. The use of on-line products creates two key issues. Firstly, who pays the cost of the license? The tutor will typically have to convince their department/school or faculty to finance the cost of any licenses as it is difficult to insist that students pay. Secondly, how do you limit the potential for cheating? Many on-line products have features to deter copying. These include:

- The use of pooling. With pooling, the tutor creates different versions (i.e. a pool) of each multiple-choice question i.e. typically five to ten alternatives. It is easy to find a number of similar questions in the various banks of questions. The software randomly selects each question from the pool so it is highly unlikely that any two students will get the same questions on the test. Some care needs to be taken to make sure the difficulty of each question in a given pool is similar.
- Randomising the order in which the questions appear on the test.
- Randomising the order of the answer options to any given multiple choice question
- Setting time limits for completion of the test. Once the student begins the on-line test, they have certain amount of time to complete the questions i.e. 30 minutes.
- The release of grades/feedback after the deadline.

Some recent research indicates that this type of approach can have a positive impact on learning. Chevalier, Dolton and Luhrmann (2017) find that graded quizzes increases the examination performance of economics students by 0.27 of a standard deviation. There is also no evidence of any displacement effects i.e. grades and pass rates in other modules are not negatively affected. However, a potential drawback is that it may encourage low quality activities i.e. those that generate surface learning. There is evidence that students believe that rote learning and memorising course content are effective strategies to perform well in multiple-choice tests (Scouler, 1998). This focus on rote learning might also crowd out other higher quality learning activities that lead to a deeper understanding of the material.

Some tutors also question the extent to which automated on-line tests can measure high order skills of analysis, synthesis and evaluation. Section 4.1 of the handbook discusses the use of multiple-choice tests in more detail.

The module requires the submission of a number of assessments that tutors grade in a relatively low cost manner

The following case briefly outlines an example

Case study 3.3.1: A log - book exercise

Some economics departments use this type of assessment in both Intermediate Microeconomics and Intermediate Macroeconomics modules. Students have to write and submit four problem sheets that contribute ten per cent of the module mark i.e. 2.5 per cent for each individual problem sheet. The tutor releases the exercise a week in advance and the students write their answers during four specified seminars. The questions on the problem sheets typically involve the completion of a set of short numerical problems and/or the representation of solutions using diagrams. Each exercise is marked on a pass/fail basis to keep the grading process as simple as possible. The tutor posts model solutions on the virtual learning environment. The marking criteria are as follows.

Pass

Students receive a pass grade for work of a good standard in terms of both quantity and quality. While there may be some errors or omissions, the answers demonstrate evidence of a good understanding of the core aspects of the material. There is also evidence of good preparation.

Fail

Students receive a fail grade for one or more of the following reasons.

- The work reflects a deep and fundamental misunderstanding of core aspects of the material.
- There is an unacceptably high frequency of mistakes and errors in the work that indicate undue carelessness and/or a complete lack of preparation.
- The amount of work completed during the session is unacceptably low, such that significant elements of the problem sheet are missing or incomplete.
- Absence from the class.

An alternative approach is to require students to complete a number of ten-minute mini-assessments at the end of seminars. The questions on each mini-assessment are very short and similar in nature to the problems on the seminar sheets. Once again, this makes them easy to mark. The highest five marks from the six mini-assessments count towards the final grade.

The module requires the submission of a number of assessments but the tutor does not grade them

Students have to complete all or the majority of the problem sheets in order to be eligible to complete other assessments i.e. the final exam. The tutor does not mark or grade the work and simply provides model answers. One obvious issue with this design is the quality of the work. How much effort will students exert on the tasks if it is not marked and graded? Checking to see if the students have submitted all

the problem sheets also involves some administrative costs. One final issue is the credibility of the threat to exclude students from other assessments i.e. is it consistent with university assessment regulations?

The module requires the submission of a number of assessments but the tutor only marks and grades a fraction of the work

In this assessment design, students have to submit a minimum number of assessments during the module i.e. the answers to 4/5 problem sheets. The tutor posts guideline solutions for each exercise on the VLE. After the deadline for the last problem sheet, the tutor randomly chooses and grades just one. This mark counts towards the final grade.

The tutor needs to take care that each of the problem sheets is of approximately the same level of difficulty. This approach may seem very unusual but in many ways mirrors that of an examination where tutors provide very limited guidance about the topics, i.e. students' understanding of some of the module content they have learnt is never measured/graded. It also encourages consistent study habits, i.e. taking each problem sheet seriously because of the random selection of the one that is graded. It may induce more consistent effort than marking all the problem sheets: students may exert less effort on a problem sheet if it is only carries a few marks.

The module requires the submission of a number of assessments that the students grade

With some practice, students may be able to mark the work of their peers effectively. It is possible for tutors to organise peer marking in seminars by providing answer guidelines and moderating a sample to check for consistency. Some research has found that peer marking by students to be as reliable as that of lecturers. Mostert and Snowball (2013) discuss the use of on-line peer assessment in a first-year macroeconomics module with over 800 students.

4. Different types question and some alternative assessments

Questions in either coursework or examination fall into two broad categories – fixed response and open/constructed response. With fixed response assessments, students choose between the possible answers provided by the tutor. Popular examples include true or false and multiple-choice questions. With open response questions, students have to construct an answer either verbally or in written form. Examples include short answer questions, essays, reports and presentations. Both fixed and open response questions are widely used on most economics programmes. Each has relative strengths and weaknesses in the way they measure and support learning and the costs they impose on tutors.

4.1 Multiple choice

4.1.1 Standard multiple-choice assessments

Standard multiple-choice questions are one of the most widely used types of assessment on economic programmes both in the UK and in other countries. They are particularly prevalent in Principles of Economics modules and are utilised in both coursework and exam. Referred to in the assessment literature as SA (single answer), NR (number of right answers) or NC (number of correct answers), the assessor has to construct a question, called a stem, and a number of alternative answers – typically four or five. One of the alternative answers is correct, while the others are distractors i.e. incorrect answers.

Correct answers receive a positive score while incorrect and unanswered questions receive a mark of zero. To calculate the total mark for the test, the tutor simply multiplies the number of correct answers by the mark per correct answer – usually a constant number.

Why are they so popular? What are the advantages?

- They enable the assessor to test for knowledge and understanding across a broad range of economic/quantitative topics.
- As a measurement tool, there are no issues with reliability/consistency. This saves on the time and effort required to ensure consistency when using other types of assessment (e.g. moderation activities).
- It reduces the costs of marking especially for tutors on modules with big student numbers.
- It is feasible to mark a large number of answer papers very quickly by tutors with no expertise in the topic area. Many universities have also invested in technology and machines to automate the marking process. This enables students to receive their grades promptly after submission of the work.

- Automated analysis of the results (i.e. scores by question) is also possible.

There are also a number of disadvantages

- Although they reduce the marking costs, effective multiple-choice questions are difficult to write and take longer to construct than many other types of assessment. Writing a stem typically involves constructing a question, a statement or incomplete statement with blanks. The tutor then has to compose the correct answer and a number of distractors. Writing effective distractors is often the most difficult and time-consuming part of the process. Haladyna and Downing (1989) provide the following guidelines:
    - When using a completion stem, avoid placing the blank at the beginning or middle of the statement.
    - Use the phrase 'all of the above' as one of the answers as sparingly as possible.
    - Use the phrase 'none of the above' as one of the answers as sparingly as possible.
    - Take care to make sure all distractors are plausible and impossible to rule out without some knowledge/understanding of the module content.
    - Use familiar yet incorrect phrases as distractors.
    - Use true statements that do not answer the question.

  It is possible to reduce the costs of setting this type of assessment by using banks of prewritten multiple-choice questions. However, tutors still need to take some time checking the quality of any imported questions as they can vary considerably. MyLab Economics and Aplia are two of the most commonly used resources in the UK and these on-line products support some of the best-selling textbooks. Another alternative is to encourage students to generate their own multiple-choice questions using PeerWise. This online repository enables students to create, answer, rate and discuss questions. To incentivise participation, tutors can use some of the best questions in the graded assessment. It is important to communicate this intention to the class.

- Students with no understanding or knowledge of the material can score marks by guessing the correct answer. With the grading system in standard NC multiple-choice assessments, students have an incentive to attempt questions, even when they have no understanding of the material. This reduces the validity of the assessment.
- Standard NC questions treat learning as if it was a dichotomous variable i.e. the student receives a mark for complete understanding (i.e. choosing the correct answer) or zero for absence of knowledge. In reality, learning and understanding is a continuous variable with students having varying degrees of knowledge and understanding and the measurement instrument should

reflect this. For example, open response assessments enable students to score some marks on a given question for demonstrating partial understanding of relevant content.

- Buckles and Siefried (2006) argue that they are not an effective way of testing for the highest levels of understanding, synthesis and evaluation. However, the authors of the Test for Understanding College Economics (TUCE) claim that multiple-choice questions can be used to measure the higher levels of Bloom's taxonomy of educational achievement.
- As discussed below, the use of multiple-choice questions may encourage students to engage in less productive learning activities i.e. surface learning and memorisation.

4.1.2 Some different types of multiple-choice assessment

The structure and grading of multiple-choice assessments can be adapted in a surprising number of ways. Some of these have been carefully trialled and tested to see if they address the limitations of using standard NC tests. The following section will discuss some of these alternatives.

Negative marking

The most widely used alternative to standard multiple-choice assessments is negative marking for incorrect answers. The main rationale for this approach is to deter guessing. One of the key issues with negative marking is the optimal size of the penalty. How large does it need to be to deter guessing?

Many assessors set the size of the penalty so that the expected score from guessing, for a candidate with no understanding of the material, is equal to the certain score of failing to answer the question. This is possible by setting the penalty equal to S/(C-1), where S is the score for a correct answer while C is the number of alternative answers in the multiple-choice questions. Therefore, where S =1 and each multiple choice test question has five possible answers, the penalty is set to 1/(5-1) = 0.25. The expected score from guessing is 0.8(-0.25) + 0.2(1) = 0 i.e. the same from missing out a question. If students are risk averse and have no understanding of the material, they should always leave the question blank, as the certain score from omitting an answer is the same as the expected score from taking a risky guess.

Negative marking does not effectively address the issue of guessing where students have partial knowledge. If they are able to identify some of the distractors, the expected score from guessing is positive[1] and the student's decision to guess now depends on their risk preferences. This same issue could also arise if some of the distractors are poorly written i.e. obviously incorrect/completely implausible. This introduces variations in test scores that are unrelated to the depth of learning i.e. it reduces validity. Students with similar understanding of the material may

differ in their willingness to answer questions as they have different attitudes towards risk.

Survey evidence also suggests that negative marking is unpopular with students. A common complaint is that it makes the assessments more stressful. This may reflect an element of loss aversion. One way to address this issue is to reward students for failing to answer questions rather than penalising incorrect answers i.e. rewarding desirable behaviour as opposed to penalising undesirable behaviour. For example, tutors could use the following grading design. Students receive one mark for a correct answer, 1/C for an unanswered question and no marks for an incorrect answer. The expected score for guessing when students have no knowledge is 0.2(1) + 0.8(0) = 0.2, the same as the certain score from failing to answer the question. This removes the loss framing, but introduces an element of grade inflation. It also fails to address the issue of partial knowledge.

Another concern with both of these approaches is that students will spend far too much time strategically thinking whether to attempt a question rather than simply focussing on the economic content of the assessment.

Elimination Testing

Elimination testing (ET) is a way of rewarding partial knowledge and so uses a more continuous measure of learning. Rather than trying to identify the correct answer, students have to indicate which of the answers they believe are incorrect. In other words, they have to identify the distractors. They can choose to eliminate up to a maximum of C-1 of the suggested answers. The following scoring system is the most widely used with ET:

- For each distractor a student correctly eliminates, they receive a mark of S/(C-1) where S is the number of marks awarded for a correct answer in a SA test.
- If they incorrectly eliminate the correct answer, they receive a penalty of –S.

Therefore where S = 4 and C=5, the student receives one mark for each distractor they correctly eliminate. Identifying all four distractors on a question scores a mark of four. However, if they eliminate two distractors and the correct answer they receive a score of -2 (+1, +1, -4).

Bradbard, Parker and Stone (2004) discuss the implementation of ET in an undergraduate macroeconomics module. Some tutors worry that negative marking will reduce the distribution of scores for the assessment, so the authors grade the work by adjusting the raw ET score to take account of the range of negative marks. For example, where S = 4 and C =5, a student's mark on a single question can range from -4 to +4. Therefore, on an assessment with 25 questions it can range from -100 to + 100. To calculate the percentage score, the authors add hundred

marks to a student's raw score and then divide by 200. For example, a student with a raw score of 60, is awarded a percentage mark of 80 [= (60 + 100)/200]. They conclude that this approach reduces the incidence of guessing and measures partial understanding in a more effective manner.

Subset selection testing

Subset selection testing (SST) is very similar to ET. However, instead of trying to identify the distractors, the students need to identify the correct answer. The approach is different from standard NC assessments as the students can choose more than one potentially correct answer i.e. they can choose a subset of answers. However, the more answers they choose, the lower the mark. If a student successfully chooses one correct answer, the result is equivalent to NC or identifying all of the distractors in ET. For each distractor included in the subset, the mark falls by S/(C-1). Therefore if S = 4 and C=5, a score of 3 is awarded if the students chooses two options that include the correct answer and a distractor (+4, -1). If the two options chosen are both distractors, the score is -2 (-1, -1). Otoyo and Bush (2018), outline a subset selection design without any negative marking.

Confidence based marking

Confidence based marking (CBM) is another alternative design that aims to deter guessing and reward partial knowledge. As with standard multiple-choice tests, students select a single option from the choice of answers or decide to omit the question. If they select an option, they have to choose a level of confidence for that answer. Most CBM schemes use a three-point scale i.e. students have to choose C = 1 (low), C = 2 (high) or C = 3(high). The confidence level chosen determines the size of the positive score if the answer is correct and the size of the negative score if it is incorrect. See an illustrative scheme in table one below.

Table one

Stated confidence level	C=1 (low)	C=2 (mid)	C=3 (high)
Mark if answer is correct	1	2	3
Mark if answer is incorrect	0	-2	-4

Gardner-Medwin and Gahan (2003) argue that tutors need to take care with the design of the scores so that they motivate the desired behaviour. For example, if the penalties for incorrect answers are -1, -2 and -3 in table 1, then it is never in the students interests to choose C=2 no matter what their level of confidence.

As with negative marking, some argue that attitudes towards risk will influence marks in a CBM scheme. For example, it may disadvantage female students who

tend to be more risk averse. However, Gardner-Medwin and Gahan (2003) find no evidence of any gender differences in the data.

Top Tip:

When introducing any alternative to traditional multiple-choice assessments it is important to provide clear instructions and plenty of practise opportunities before the students take the test that is graded.

---

[1] If a student can correctly identify three distractors and believes that both remaining answers are equally likely to be correct then the expected score from guessing is 0.5(1) + 0.5(-0.25) = 0.5 – 0.125 = 0.375.

4.2 Short answer assessments

Short answers questions are a type of open response assessment. The following are some examples:

Write short briefing notes on the following, explaining each concept and its significance for macroeconomic policy:

1. Inflation bias
2. The dynamic aggregate demand curve
3. Credit-constrained households
4. The political business cycle
5. Public-sector primary surplus

Write short briefing notes on the following threshold concepts explaining each concept and its real-world significance.

1. Markets may fail to meet social objectives
2. Rational decision making involves choice at the margin
3. People's actions depend on their expectations
4. Elasticity of a variable to a change in a determinant
5. The distinction between nominal and real values

The use of these types of question enables tutors to measure students' understanding of a broader area of the curriculum. It is also possible to measure and reward partial knowledge. Short answer questions are one of the easier types of assessment to construct and are a useful way to develop generic skills such as writing concisely, identifying key issues and communicating to different audiences. For example, they might mirror the short briefing style used by professional economists or an abstract/executive summary written by academics.

One potential drawback they have in common with multiple-choice assessments is erroneously signalling that only superficial learning of module content is required. They may be an effective way to develop and test a student's ability to apply economic theory but some tutors question their suitability for measuring higher order skills such as synthesis and evaluation. If the question is broken down into small parts, there may be a tendency to award all the marks (3/3) or none at all (0/3) and the assessment faces the same measurement issues as standard multiple-choice questions. Ensuring consistency also requires some moderation activities.

Top Tip:

It is important to make sure students and tutors have shared expectations about the appropriate length/depth of short answer questions. Sometimes, students falsely believe an appropriate 'short answer' is one or two sentences when the tutor anticipates half to a full page of writing to address the question in enough depth.

4.3 Extended open response assessments

Much longer open-response assessments, such as essays, are popular as many tutors view them as a more effective way of developing and measuring higher order skills. For example, to construct and sustain an academic argument in longer written answers, students have to internalise and develop a deeper understanding of economic theories and concepts. Walstad (2006) argues that:

"An essay question challenges students to select, organise, and integrate economics material to construct a response – all features of synthesis. An essay question is also better for testing complex achievement related to the application of concepts, analysis of problems, or evaluation of decisions. This demonstration of complex achievement and synthesis is said to be of such importance as a learning objective that it is used to justify the extra time and energy required by the instructor for grading the essays."

Assessors can write essay questions in a number of different ways. The following are some of the more common styles with examples shown in italics.

- Reproduce and explain relevant economic theory.

Explain the theory of perfect competition.

- Compare and contrast two or more different economic theories.

  With reference to the lottery choice experiments that you played 'in class', assess the major differences between expected utility theory and prospect theory.

- Apply economic theory to a real-world issue and or policy.

  Using the concept of externalities, discuss the economic rationale for imposing a per unit tax (i.e. an excise duty) on alcoholic drinks.

  Using the AD/AS framework, discuss the possible shot-term and longer-term adjustment of an economy to a negative demand-side shock.

- Summarise relevant empirical evidence and judge the extent to which it supports the predictions of economic theory.

  Evaluate the argument that as alternative methods of raising the welfare of a target group of consumers, unconditional cash transfers are more effective than in-kind transfers. To what extent does the evidence suggest that the labelling of these transfers influences the way they are spent?

  To what extent does the evidence suggest that Giffen goods exist in reality?

- Use economic theory/evidence to appraise a point of view, opinion or assertion.

  "The predicted outcomes of a monopolistically competitive market are more efficient than those of a perfectly competitive market." Discuss this assertion.

  Evaluate the argument that in highly financialised economies like the UK, the balance sheets of economic agents are an important source of economic volatility.

  Discuss the argument that supply-side factors are the sole determinant of the economy's potential output.

Some issues with extended writing

Tutors face a number of issues when using extended writing assessments.

- Should the essay/report be broken down into sections with marks clearly allocated to each section? For example:

  Compare and contrast the economic model of perfect competition with that of monopolistic competition paying particular attention to:

a. *their characteristic features and assumptions (20 marks)*
b. *the nature of the long-run market adjustment (40 marks)*
c. *their implications for economic efficiency (40 marks)*

This may help to provide effective guidance on how to structure the work and develop a shared understanding of what is important. However, breaking down questions in this way can also send a very prescriptive message to students about what to write and so deter deeper learning. It also constrains the ability of markers to reward high quality answers to particular sections of a question.

- Rather than trying to develop an internalised and deeper understanding of economic theory, students simple cut and paste material from various sources on the internet. This may lead to the inclusion of maths/diagrams, which are either irrelevant or not adapted to the issue referred to in a particular essay question.
- It increases the time it takes to mark work and provide feedback. This may become unmanageable for modules with over a hundred students.
- Grading the quality of essays involves the tutor making subjective judgements. Factors such as the halo effect are more likely to influence the marking. Therefore, more effort is required to ensure both intra and inter-marker consistency.
- Tutors need to consider the levels of support they are willing and able to offer students during the writing process. One particular issue is feedback on draft copies of work. Many students expect this level of support as it is common in pre-university education. Therefore, tutors need to communicate the support they offer before releasing the work to manage expectations. If assessors provide feedback on drafts, it should focus broadly on what the student needs to do to improve the work rather than specifically correcting the work. There is often a fine line between these two types of comment. It is also not advisable to provide any indication of the grade as this leads to more complaints. For example, a number of students will claim they have carried out all of the suggested improvements but not received a significantly higher grade.

Tip: To keep marking at a manageable level the essay can have two-stages. The first stage is the submission of the draft copy that receives feedback but no indication of the grade. The final copy only receives a grade and no feedback. Another alternative is to get the students to provide feedback on each others work in a peer feedback activity. Mostert and Snowball (2013) discuss the use of this type of activity in a first year macroeconomics module with over 800 students.

Case study 4.31: Economics in the news[1]

This case discusses an alternative way of designing a written assessment. This particular example is from a course on Industrial Organisation but it could easily be adapted to other topic areas in economics.

The assignment has a number of elements. Firstly, students have to select and write a 500-word academic literature review on any topic covered in the module i.e.

innovation, competition policy. The relatively low word limit helps to develop concise writing skills. For the second part of the assignment, they have to find three articles in The Economist or a broadsheet newspaper relating to their chosen topic area. Importantly, the publication date for the articles must be in the same term/semester as the teaching of module. Students have to write 500-words on each article and discuss the extent to which they either support or contradict the academic literature on the topic. For the final part of the assignment, they have to write an appendix that lists ten further articles that relate to any topic on the module. Once again, the publication date for the chosen articles must be in the same semester as the module.

Some advantages with this assessment design.

- It motivates students to keep-up to date with economics news and see the real world relevance of the technical material covered in the module. This can be very useful interview preparation for a graduate job.
- It encourages more consistent studying during the module as opposed to very intensive bursts of effort just prior to a deadline. At a minimum, students need to be reading and selecting relevant articles throughout the term. To reinforce the consistency of study effort, tutors could spend a few minutes at the beginning of each seminar asking the class for examples of articles they have found in the preceding week.

Some students can be unsure and nervous about this type of coursework as it is different to what they have encountered previously in their studies. For this reason, it is advisable to use some contact time to discuss anonymised examples of work from previous years. Some tutors are nervous about discussing exemplars as they fear students will simply copy them in their own work. However, the requirement for articles from the current academic year reduces the likelihood of this occurring.

---

[1] For more detail see Elliott and Balasubramanyam (2016)

4.4 Some innovative types of coursework

The following cases outline some alternatives to the standard methods of written assessment.

Case study 4.4.1: Using class debates[1]

The use of debates is a suitable assessment design for a wide range of topics such as (a) the Bank of England's decisions on interest rates (b) the UK's decision to leave the European Union (c) A World Economic Forum on policies to combat

climate change and (d) a firm's decision to enter an overseas market. This particular case is from a course on competition policy.

Students self-select into groups of 4-5 and choose a competition case they want to investigate from a list prepared by the tutor. Two groups choose each case. One represents the competition authority (i.e. the European Commission, Competition and Market Authority) while the other represents the business under investigation (i.e. Microsoft; Google; Intel; Qualcomm).

Before the debate, it is important to hold an initial meeting with both groups to set the scene and highlight some of the key features of the chosen case. Then, about a week before the debate is scheduled, the tutor meets with each group separately to check their arguments and answer any queries. These meetings play a key role in ensuring the quality of the presentations.

The actual debate is organised in the following manner:

- For the first 15 minutes, the group representing the competition authority make their arguments. The group representing the firm has the next 15 minutes to respond.
- For the next 10 minutes, they have the opportunity to cross-examine each other. The tutor also asks some of their own questions as well as any from the other students in the audience.
- Each group is given a final minute to summarise their key arguments
- Finally, based on the debate, the audience vote on what they believe should be the outcome of the case. This helps to maintain engagement during the session.

The tutor makes it clear that in some cases one group has a much harder position to defend and argue. Therefore, the grade does not depend on the final audience vote. Instead, the assessment criteria includes factors such as the presentation of the arguments, links to economic theory, cross-examination and response to questions and summary of arguments.

When the debate format replaced a more traditional group presentation, student engagement improved both during the preparation and delivery stages. The design seems to tap into the competitive desire of students to outperform their peers and the feedback has always been positive. From the tutors viewpoint, assessing and grading is far more enjoyable than sitting through numerous standard presentations.

Case study 4.4.2: Using videos

The increasing number of students, who have high quality video cameras on their mobile phones and tablets, makes this type of assessment far easier to implement

than in the past. The following example is from a final year option module on behavioural economics.

The assessment asks the students to apply Behavioural Economics to any real world issue or problem. The module leader supplies some possible topics but students are free to choose as long as the tutor agrees that it is a suitable area. The assessment brief clearly states that all students must contribute. If a member of the group does not appear in the video, he/she has to provide a short-note outlining their contribution to the work. The guidelines also stress that the tutor will grade the video on its economic content rather than the quality of the recording or editing. The students can also use any type of video recording equipment.

Each group has to produce a 3-minute video. The tutor chose this duration, as it is the same as the requirement for the Royal Economic Society Undergraduate Video Competition. The submission is via the virtual learning environment or YouTube.

Some advantages over traditional in-class group presentations.

- It frees up class contact time to do other activities.
- Students no longer have to sit through numerous presentations of variable quality. They can watch the videos submitted by the other groups, if they are interested in the topic.
- It gives tutors greater flexibility over when they watch and grade the presentations. This could help to reduce problems of marker fatigue.
- The quality of the presentations tends to be much higher than those completed in-class. One potential reason for this is that the format helps to remove some of the anxiety of presenting directly in front of people. This allows students to focus on the economics content.

Top Tip:

The Royal Economic Society (RES) often runs an undergraduate video competition. Try to follow the guidelines in the module assessment so that students can easily submit their videos to this competition.

---

[1] For more detail see Olczak (2019)

5. Improving the impact of feedback

As previously discussed, one important way assessment can support learning is through effective communication between the tutor and student about the quality of work. Numerous studies have found that feedback has a significant impact on learning (Black and William, 1998; Hattie and Timperley, 2007; Kluger and DeNisi, 1996). However, there is considerable variability in the results. Given its potential importance, why might students fail to respond to feedback? How can we increase the likelihood that they will engage with and act upon the guidance tutors provide? The following section discusses a number of these issues.

It arrives too late

Students are more likely to engage with feedback if they receive it while the process of researching and writing the assessment is still fresh in their minds. Unfortunately writing detailed comments on hundreds of assignments can take weeks. By the time marking is complete and the feedback returned, many students have started studying for subsequent assignments. The comments may no longer seem relevant as they focus on their next assessment. There are two different ways of providing feedback quickly – even on very large modules.

- Provide feedback after reading a sample of assignments.

Instead of marking all the students' work before providing feedback, read a sample of the assignments in the first couple of days after the deadline. Identify any common weaknesses and either discuss these in the next class or post announcements/handouts on the virtual learning environment.

- Provide feedback before reading any assignments!

Tutors can often predict/anticipate common mistakes or weaknesses in students' work before they have marked a single assignment. Rather than keeping this information private, produce a handout and discuss these anticipated weaknesses with the students in the first class following the deadline date. It is also useful to spend some of this contact time describing some of the key features of a good answer.

Although imperfect and rather generic, feedback provided in the first few days following a deadline, may have a stronger impact on some students than more personalised and detailed feedback provided at a later date.

It discourages and demotivates

One purpose of feedback is to motivate students to take appropriate actions to deepen their learning. If the comments only discuss weaknesses and the language

is harsh/judgemental, it can have the opposite result. When marking, it is easy to forget the emotional responses people feel when reading comments on work in which they have invested a large amount of their own time and effort. As well as having a negative impact on self-confidence and motivation, students are likely to ignore feedback if it is overly negative. Always try to find something positive to say otherwise there is a danger the students will take no notice of any of the comments.

The impact of releasing marks/grades

There is evidence that once students see their marks/grades they are more likely to ignore potentially useful feedback. There are a number of possible explanations. For example, those students who receive high grades may believe they have mastered the topic and so do not need to read any of the comments. Those who receive low grades may feel they never want to look at or engage with the assignment ever again. One way to address this issue is to release feedback before the marks. Students then have to provide an estimate of the grade based on the comments. They are encouraged to compare the feedback with their peers and a small grade incentive for accuracy is a useful way to encourage them to take it seriously.

Top Tip:

A useful incentive scheme operates in the following way. If the estimated mark from the feedback is within five percentage points either above or below the final mark awarded by the tutor, the student receives a bonus of five percentage points. To avoid any incentives to game the system, add the five percentage points to the mark awarded by the tutor – not the estimated mark provided by the students.

It seems irrelevant for future assessments

In the research literature there is evidence that some students ignore feedback because they believe it is specific to that particular assignment and provides no guidance on how to improve their future work. Many lecturers write feedback comments as if the students have submitted a draft copy of the work for a later resubmission. How can we avoid doing this? Some comments relate to the academic content of the assignment such as the choice/explanation of economic theory and its application to any issues raised in the question. When writing these types of comments, it is important to highlight cases where a good understanding of this same academic content is required for students to perform well in subsequent assessments. For example

> *"You need to gain a deeper and more thorough understanding of expected utility theory if you wish to improve your performance in the final examination"*

Other comments relate to more generic skills development such as the structure of the answer, the balance of material, the quality of written communication and the ability to develop arguments in a logical manner. It is easier for students to see the relevance of improving these skills for future assignments but always signpost and make it as clear as possible.

Perceptions about the usefulness of all types of feedback are greatest when provided on draft versions of work. Some issues with marking drafts are considered in section 4.3.

It does not clarify the size of any weaknesses or gaps in understanding

Traditional written feedback can be effective at identifying gaps. For example, commonly used comments include

“The assignment lacks clarity and logical coherence.”

“There is not enough critical analysis.”

“Some concepts are not explained in enough detail”

“The answer did not focus on the question.”

It is far more difficult to explain the size of any gaps. For example, the second comment above identifies that there is not enough critical analysis but says nothing about the level required to achieve a particular grade. One way to address this issue is to show students concrete examples of work that demonstrate the standard or the skill at an appropriate level. These are post-submission as opposed to pre-submission exemplars. For example, when marking assessments, copy samples of answers that illustrate good performance on some aspect or aspects of the assessment criteria. Distribute these answers in class or post on the VLE. Referring to these exemplars in the written feedback can also save time by reducing the quantity of comments.

The use of post submission exemplars can also play a very useful role when staff face students who appear disinterested in constructive feedback and just want to know why they received a mark below the one they believe they deserve. Spending a few minutes getting these students to compare their own work with examples of high quality exemplars is an effective and efficient way of dealing with these difficult situations.

Top Tip:

Rather than providing complete versions of the post submission exemplars, copy a particular page or highlight a paragraph that is a good example of some element of

the assessment criteria. In the feedback, include a sentence along the following lines – "For an example of a piece of work that demonstrates excellent critical analysis see the highlighted section on exemplar A"

It does not explain how to improve

This is the trickiest part of the feedback process. It is difficult to specify/outline exactly what the student needs to do or what future actions they can take to improve. Comments such as 'you need to work harder' are unlikely to have an impact. Some alternatives include:

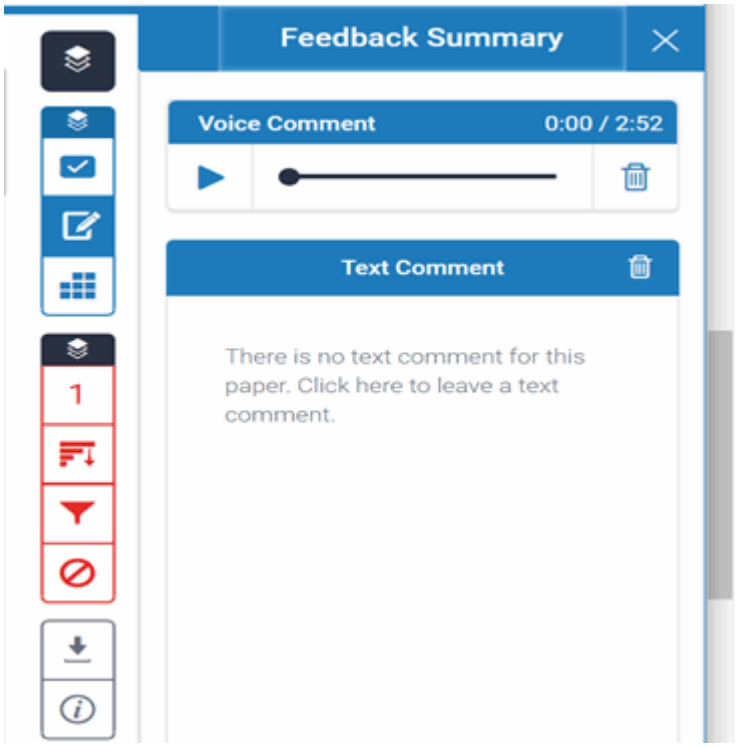
"Read lots of different examples of other assignments that received a high grade. Compare them against your own and try to identify their particular strengths and areas you need to work on to obtain higher marks in future assessments."

> *"In the future, try to read through your work more carefully and amend any errors before handing it in."*
>
> *"Book an appointment with the support centre x at the university to receive extra support."*
>
> *"Try to attempt more practice questions."*

"Go back and read chapter x again in the textbook and try to gain a better understanding of theory y."

Case study 5.1: Using audio feedback

GradeMark in Turnitin is perhaps the most common way for tutors to provide written feedback comments on coursework – see the screen shot. Just above the box for typing in text comments is a general feedback recorder i.e. see the 'Voice Comment'.



It is very straightforward to record comments using this function using headphones or the built in microphone on a PC/tablet computer. The recording can be paused at any time while the tutor reads the next section of the coursework. Although it is very easy to use, one major limitation is a three-minute limit on the duration of the audio comments. It is also impossible to edit files. If the tutor is unhappy with any of the feedback, they have to re-record all three minutes. Other software such as Audacity provide greater flexibility but involve more time costs i.e. creating the files and posting the results.

Research suggests that audio feedback is both more efficient and effective than written feedback. For example, in a simple experiment Lunt and Curran (2010) compared the effort levels of producing the same feedback comments in three different ways. On average, it took tutors three minutes to type, four minutes to hand write and forty seconds to record the comments in an audio file.

When used on a large course in intermediate microeconomics, students thought that vocal explanations conveyed meaning more effectively than written explanations. In particular, they found the comments more detailed, supportive and personalised. Some research suggests that intonation, inflection and tone in the audio comments increases the likelihood that students will respond to critical/developmental feedback. Another potential advantage is the impact on marker fatigue. A number of tutors find recording audio feedback comments over an extended period of marking less tiring than writing feedback comments.

6. Assessment of learning – Measurement Issues

One of the key functions of assessment is the measurement of learning. When marking coursework and exams, tutors make judgements about the extent to which students demonstrate (a) an in-depth understanding of the module content and (b) achievement of the learning outcomes. This typically involves the awarding of marks, grades and certification of achievements. Employers and other educational institutions frequently use this information as part of the selection process for jobs and places on postgraduate courses. This makes students acutely aware of its importance. There is a tendency for many tutors, as well as students, to focus solely on the measurement function at the expense of all the other objectives.

6.1 Validity

The measurement function raises two important issues – validity and reliability. In order to understand validity it is useful to consider the following questions. Are the current methods of assessment for the module appropriate and effective measurement tools? Do they indicate the extent to which students have met the modules learning outcomes? Put simply, do they measure what they purport to measure? Validity of assessment focuses on this issue. Why might assessment lack validity? There are two key reasons.

a. The methods of assessment fail to capture one or more of the learning outcomes. This is most likely to occur where learning outcomes refer to specific theories from the module's indicative content. For example:

'By the end of this module, successful students will be able to explain how to apply elementary game theory to economic models of firm behaviour'

The problem with writing learning outcomes in this way is that assessment does not typically test students' understanding of the entire content of the module. Coursework elements often focus on detailed investigations of a single topic while examinations include questions that cover part, but not all, of the syllabus. If an exam does include questions on the entire module content, they can typically only test a relatively superficial knowledge of each topic. In exams where students need to demonstrate a greater depth of understanding, the design includes some choice over the questions they answer. Therefore, it is usually possible for students to achieve a pass grade without demonstrating any knowledge of some areas of the syllabus i.e. 'explain how to apply elementary game theory'. For this reason, it is advisable to write learning outcomes that refer to economic theory more broadly. For example:

'By the end of this module, successful students will be able to discuss and apply a range of economic theories at an intermediate level'.

Some learning outcomes refer to generic skills. For example,

> *'By the end of this module, successful students will be able to communicate economic principles in a clear and precise manner through written work'*

If the assessment for the module is only by examination then perhaps the words '*in time-constrained conditions'* needs to be added to the end of the previous learning outcome.

Top Tips:

When you first take over a new or existing module, carefully read the existing learning outcomes. Ask yourself if they clearly articulate the learning you have in mind for the module. If you do not think this is the case, then change them.

Always keep the module learning outcomes in front of you when writing assessments. If you want to change the assessment for a module, make sure the new design is still an appropriate way to measure the learning outcomes. If this is not the case, you will need to change the learning outcomes.

b. Assessment may also lack validity where factors other than those referred to in the learning outcomes influence the grade – i.e. factors other than the students' understanding of relevant economic theory and/or demonstration of generic skills. For example, scores on some methods of assessment, such as multiple-choice tests, may vary because of luck and risk preferences (i.e. lucky guesses) rather than the underlying knowledge of the module content. (There is more discussion on this topic in section 4 of the handbook). The wording of questions can also create problems. Even very experienced tutors, find it difficult to write assessments where there is no ambiguity over precise meaning. This lack of clarity leads to some students achieving higher marks because they are lucky enough to infer the same meaning as the tutor rather than any deeper understanding of the course material.

Top Tips:

Always remember that from your students' perspective, the most important words you write in any module are the assessment questions. Make sure you give them enough time, care and attention.

It is very difficult to write assessment questions and then instantly recognise any ambiguity in their meaning. Always re-read assessment questions a number of days after writing them. Also, try to encourage as many of your colleagues as possible to read and comment on the clarity of your questions.

6.2 Reliability and Consistency

For the measurement function of assessment to work effectively, grading must be consistent both between different assessors (inter-marker reliability) and by the same assessor (intra-marker reliability). To achieve intra-marker reliability, the same assessor must award the same grade to different pieces of work of very similar quality. To achieve inter-marker reliability, different assessors must award the same grade to different piece of work of similar quality.

6.2.1 Intra-marker reliability

Sadler (1989) argues that tutors'

“Conceptions of quality are typically held, largely in unarticulated form, inside their heads as tacit knowledge”. (Sadler, 1989, p.54)

One important issue is whether the application of these ‘inner standards’ remains constant during the marking process or whether they are subject to a number of biases. For example, are judgements influenced by (a) comparisons with the quality of previously graded work (b) marker fatigue (c) changes in mood?

Contrast or sequential effects

This is where the grade awarded to any given assignment/exam script is a function of the quality of the previous assignments/exam scripts. For example, a tutor may grade an average piece of work far more harshly if they have just read a block of excellent answers as opposed to a block of weak answers. Higher quality work may also receive a higher grade when it follows a block of lower quality work as opposed to a number of outstanding pieces of work. A number of research papers (Hughes, Keeling, and Tuck, 1980; Daly and Dickson-Markman, 1982; Yeates, Moreau and Evra, 2015) report evidence of these effects. Hughes, Keeling, and Tuck (1980) found (a) the impact was stronger for average quality assessments and (b) the effect declined with the number marked. Perhaps more worryingly, Yeates, O’Neil, Mann and Evra (2013) found that assessors lacked awareness of their susceptibility to this bias.

Assessor fatigue

After a prolonged period of marking, tutors may start to read answers less carefully and fail to notice important strengths and weaknesses with the work. This leads to a smaller spread of grades as higher quality assessments receive lower marks than those of a similar quality read earlier in the day. The reverse is true for lower quality work i.e. the tutor awards higher marks. There is a danger that different pieces of work of varying quality all start receiving a mark somewhere between 58% and 63%.

Mood effects

This is where the application of the inner standard varies from one day to another based on the general mood of the marker. If the marker feels happy, they may award higher grades to similar pieces of work than when they are feeling down.

Top Tips:

Briefly read a number of the assignments to get an initial sense of the types of answers before grading and providing feedback.

When the marking is completed, sort the assignments/exam scripts in rank order. If the assignments have been submitted electronically using Turnitin, this can easily be done by clicking on the grade column. Once sorted, read through some assignments/exam answers you have awarded the same grade and check for consistency.

Marking exam paper by question, as opposed to the whole script, may help to improve consistency.

6.2.2 Inter-marker reliability

This is a major issue especially on large modules where a number of different tutors mark and grade coursework and exam scripts. Do they have similar or different perceptions of the standard required to achieve particular grades? Are some markers systematically more lenient than others? There could also be a 'halo effect' where the quality of one aspect of the assessment overly influences the tutor's judgements on other aspects of the assessment. Examples could include the impact of (a) high standards of presentation and (b) the quality of the opening paragraph. The strength of these biases is likely to vary between different markers.

In an ideal world, whoever marks the work should have no impact on the grade awarded. Unfortunately, research evidence suggests that there are wide variations in the grades awarded by different tutors for work of similar quality (Baume, Yorke and Coffey, 2004; O'Hagan and Wigglesworth, 2014). In an attempt to deal with this issue, many universities expect module leaders to provide written guidance that clearly outlines the criteria used to grade coursework. For example:

- A detailed assessment criteria for each piece of work with grade descriptors for each criteria
- Marking schemes
- Model answers

Although the research evidence suggests that guidance helps to reduce levels of inconsistency, it remains a significant issue. For example, Bloxham, den-Outer, Hudson and Price (2016) studied the consistency of grading in four disciplines -

psychology, nursing, chemistry and history. Six highly experienced markers, in each discipline, graded five different answers to the same question with the same detailed assessment criteria. In nine cases, different tutors ranked the same assignment as either the best or the worst answer! Only one assignment received the same ranking. Why is the provision of detailed assessment criteria not enough to ensure consistency between different markers? There are a number of reasons:

- Some tutors may have very established ideas about the appropriate standards and criteria i.e. what they are looking for from a good piece of work. If these differ from the published guidance then they may:
    - Simply ignore the whole criteria and judge the work against their own unarticulated 'inner standards'
    - Ignore some of the published criteria and judge against some additional criteria from their 'inner standards'
- Some published criterion are often extremely broad and effectively require the application of sub criteria that are unpublished. Once again, the application of these unpublished sub-criteria may vary between different tutors depending on their inner standards.
- The precise understanding and application of a particular criterion may vary between different tutors. For example, does 'critical analysis' mean the same thing to all markers?
- Tutors may agree on the meaning of specific criterion but not the standard that students need to demonstrate for a particular grade.
- The weighting of each of the criteria may vary between different tutors.
- Assessors may judge the overall quality of the assessment before working backwards and applying arbitrarily generated marks for each of the published criteria.

How to improve inter-marker consistency

- The module leader provides other assessors with examples of the work they have already marked across a range of different grades.
- Tutors mark a small sample of the work and get this moderated by the module leader before continuing with the rest of their marking.
- Module leaders and other assessors grade a sample of the work together and discuss the rationale behind their grading.

Although this last approach maybe the most time consuming it is also the most effective. The Quality Assurance Agency (QAA) recommends the use of:

"Practices which promote and support consistency of marking by and between staff, including dialogues which enable a shared understanding of standards" (QAA Quality code, chap 6, p13)

Bloxham, Hughes and Adie (2016) recommend that the discussions of assessment tasks and the appropriate standard should actually take place before the teaching of the module begins.

6.2.3 The trade-off between reliability and validity

Improving the reliability and validity of assessment will often involve trade-offs. For example, increasing the proportion of quantitative/technical questions and reducing the proportion of discursive questions will probably improve both intra and inter-marker consistency. However, it may reduce the validity of the assessment i.e. the development of critical thinking and evaluative skills.

For example, the Quality Assurance Agency Benchmark Statement for economics suggests that the main aims of a degree programme in economics should include the following:

- to stimulate students intellectually through the study of economics and to lead them to appreciate its application to a range of problems and its relevance in a variety of contexts
- to develop in students an ability to interpret real world economic events and critically assess a range of types of evidence
- to foster an understanding of alternative approaches to the analysis of economic phenomena
- to equip students with appropriate tools of analysis to tackle issues and problems of economic policy

It is questionable whether it is possible to test these learning outcomes without a significant discursive element to a number of the assessments on the programme.

7. Summary

Assessment is complex and plays a crucial role in determining how much students learn. Given its importance, it is worrying that surveys report such low levels of satisfaction with this aspect of higher education. One potential explanation for this finding is a tendency for tutors to concentrate on the content and delivery of a module.

This handbook chapter discusses some implications of assessment design within in a module for the quantity, consistency and quality of learning activities undertaken by students. It considers alternative styles of assessment question and outlines some innovative types of assessment i.e. the use of class debates and videos. It also examines some ways of increasing the likelihood of students engaging with and acting upon feedback. The final section of the handbook discusses some measurement issues, and includes some advice on how to improve both the validity and reliability of assessments.

Given the multidimensional nature of assessment, it is impossible for one handbook chapter to discuss all the key issues in detail. In other handbook chapters, Cortinhas (2017) outlines some different ways of deterring plagiarism, Smith (2016) discusses undergraduate dissertations and Watkins (2005) examines the use of group work.

In the wider literature, Cook (2016) examines some innovative ways of assessing students on statistics and econometrics modules. Grogan (2017) and Green, Bean and Peterson (2013) discuss some different ways of using written coursework in economics modules.

Another issue not considered in this chapter is how to develop a shared understanding between tutors and students of what 'good' work looks like before the final submission. Guest and Riegler (2017) examine some evidence on the relative inaccuracy of economics students' self-evaluation estimates. Their findings suggest that for many students, a shared understanding of standards remains elusive. Wilson (2015) outlines some ways of better preparing students for assessment in economics, while Guest (2019) discusses the use of peer review.

Assessment and feedback remains one of the most difficult aspects of teaching. Perhaps we all need to spend a little more time reflecting on our current practice and consider some of the available alternatives.

References

Baume, D. Yorke, M and Coffey, M. (2004) What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education*, 29:4, 451-477. DOI: 10.1080/0260293310001689037

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–75. DOI: 10.1080/0969595980050102

Bloxham, S. Hughes, C. and Adie, L. (2016) What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices, *Assessment & Evaluation in Higher Education*, 41:4. DOI: 10.1080/02602938.2015.1039932

Bond, E., Bodger, O, Skibinski, D, Jones D., Restall, C., Dudley E. & van Keulen. (2013) Negatively-Marked MCQ Assessments That Reward Partial Knowledge Do Not Introduce Gender Bias Yet Increase Student Performance and Satisfaction and Reduce Anxiety, *PLoS ONE*, Vol 8, No 2. DOI: 10.1371/journal.pone.0055956

Bradbard, D., Parker, D. and Stone, G. (2004) An alternate multiple-choice scoring procedure in a macroeconomics course, *Decision Science Journal of Innovative Education*, Vol 2, pp 11-26. DOI: 10.1111/j.0011-7315.2004.00016.x

Buckles, S. & Siegfried, J. (2006) Using Multiple-Choice Questions to Evaluate In-Depth Learning of Economics, *The Journal of Economic Education*, Vol 37:1, pp 48-57. DOI: 10.3200/JECE.37.1.48-57

Chevalier, A. Dolton, P. and Luehrmann, M. (2017) Make it count: Students incentives and effort, *Journal of the Royal Statistical Society*, Vol 181, 323-349. DOI: 10.1111/rssa.12278

Cook, S. (2016) Modern econometrics: Structuring delivery and assessment, *Cogent Economics & Finance*, Vol. 4, (1) DOI: 10.1080/23322039.2016.1152705

Cortinhas, C. (2017) Detection and Prevention of Plagiarism in Higher Education, in *The Handbook for Economics Lecturers*.

Daly, J. A. & Dickson-Markman, F. (1982) Contrast effects in evaluating essays. *Journal of Educational Measurement*, v19 n4 p309-315. JSTOR: 1435003

Elliott, C. & Balasubramanyam, V. (2016) Assessing students: Real-world analyses underpinned by economic theory, *Cogent Economics & Finance*, Vol.4, (1). DOI: 10.1080/23322039.2016.1151171

Gardner-Medwin A. & Gahan M. (2003) Formative and Summative Confidence-Based Assessment. *Proceedings of the 7th International CAA Conference*, Loughborough University, UK, pp. 147-155

Gibbs, G. and Simpson, C. (2005) Conditions Under Which Assessment Supports Students' Learning, *Learning and Teaching in Higher Education* Vol (1). pp. 3-31

Green, G., Bean, J. and Peterson, D. (2013). Deep learning in intermediate microeconomics: Using scaffolding assignments to teach theory and promote transfer. *Journal of Economic Education*, Vol. 44 (2), pp. 142–57. DOI: 10.1080/00220485.2013.770338

Grogan (2017) Will this be on the test? How exam structure affects perceptions of innovative assignments in a masters of science microeconomics course, *International Review of Economics Education*, Vol 26, pp.1-8. DOI: 10.1016/j.iree.2017.06.001

Guest (2019) 'Providing Effective Feedback', in Chapman, C. Daniels, K., Elliott, C. and Finlay, S. (ed.) *How to Teach in a Business School*, Aston Business School, Aston University, Edward Elgar Publishing

Guest, J. and Riegler, R. (2017). Learning by doing: do economics students self-evaluation skills improve? *International Review of Economics Education*, 24, pp. 50-64

Haladyna, T. & Downing, S. (1989) A Taxonomy of Multiple-Choice Item-Writing Rules, *Applied Measurement in Education*, Vol. 2:1, 37-50. DOI: 10.1207/s15324818ame0201_3

Hattie, J. A., Biggs, J., & Purdie, N. (1996). Effects of learning skills intervention on student learning: A meta-analysis. *Review of Research in Education*, 66, 99–136. DOI: 10.3102/00346543066002099

Hattie, J. & Timperley H. (2007) Power of Feedback *Review of Educational Research* , Vol. 77, No. 1, pp. 81–112 DOI: 10.3102/003465430298487

Hughes, D. C., Keeling, B., & Tuck, B.F. (1980a) Essay marking and the context problem. *Educational Research*, v22 n2 p147-148. DOI: 10.1080/0013188800220207

Hughes, D. C., Keeling, B., & Tuck, B.F. (1980b) The influence of context position and scoring method on essay scoring, *Journal of Educational Measurement*, v17 p131-135. DOI: 10.1111/j.1745-3984.1980.tb00821.x

Hughes, D. C., Keeling, B., & Tuck, B.F. (1983) The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational & Psychological Measurement*, v34. DOI: 10.1177/001316448304300413

Kluger, A. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. DOI: 10.1037/0033-2909.119.2.254

Laughton, D. (2013). Using audio feedback to enhance assessment practice - an evaluation of student and tutor experiences. *Student Engagement and Experience Journal*, **2** (2). DOI: 10.7190/seej.v2i2.68

Lunt T, and Curran J. (2010) Are you listening please? The advantages of electronic audio feedback compared to written feedback. *Assessment & Evaluation in Higher Education*. Vol. 35(7):759–69. DOI: 10.1080/02602930902977772

Miller, C. & Parlett, M. (1974) *Up to the Mark: a study of the examination game*, Guildford: Society for Research into Higher Education.

Mostert, M. and Snowball, J. (2013). Where angels fear to tread: On-line peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education, 38*(6), pp.674–686. DOI: 10.1080/02602938.2012.683770

Olczak (2019). 'How to invigorate group presentations?', in Chapman, C. Daniels, K., Elliott, C. and Finlay, S. (ed.) *How to Teach in a Business School*, Aston Business School, Aston University, Edward Elgar Publishing

O'Hagan, S. R. & Wigglesworth, G. (2014) Who's marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context, *Studies in Higher Education*, Studies in Higher Education, 40:9, 1729-1747. DOI: 10.1080/03075079.2014.896890

Otoyo, L. and Bush, M. (2018) Addressing the Shortcomings of Traditional Multiple-Choice Tests: Subset Selection Without Mark Deductions, *Practical Assessment, Research & Evaluation*, Vol. 23:18. DOI: 10.7275/k1c0-pk31

Sadler, R. D., 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), pp.119-144. DOI: 10.1007/BF00117714

Scouler, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay, *Higher Education*, Volume 35, Issue 4, pp 453–472. JSTOR 3448270

Smith, P. (2016). Undergraduate Dissertations in Economics, in *The Handbook for Economics Lecturers*.

Snyder, B. (1971). *The Hidden Curriculum*, Cambridge, MA: MIT Press.

Thomas, L., Hockings, C., Ottaway, J., and Jones, R. (2015). Independent Learning: student perspectives and experiences. York: HE Academy

Yeates, P., Moreau, M., and Eva, K (2015). Are examiners' judgments in OSCE-style assessments influenced by contrast effects? *Academic Medicine*, Vol. 90, No. 7. DOI: 10.1097/ACM.0000000000000650

Yeates, P., O'Neil, Mann, K., and Eva, K (2013). You're certainly relatively competent: assessor bias due to recent experiences. *Medical Education*, 47(9):910-22. DOI: 10.1111/medu.12254

Walstad, W. (2006) 'Testing for depth of understanding in economics using essay questions', *Journal of Economic Education*, Vol. 37(1), pp. 38–47. JSTOR 30042685

Watkins, R. (2005) Group work and Assessment, in *The Handbook for Economics Lecturers*.

Wilson, C. (2015) Better Preparing Students for Assessment: Marking Criteria, Mock Assessments and Peer-Feedback, *Case Study, Economics Network*

Vanderoost J., Janssen R., Eggermont J., Callens R., De Laet, T. (2018) Elimination testing with adapted scoring reduces guessing and anxiety in multiple-choice assessments, but does not increase grade average in comparison with negative marking. *PLOS ONE* 13(10). DOI: 10.1371/journal.pone.0203931