Quality Assurance and Procedural Security in Learning and Examinations at Oslo University College

Mark Burgess

January 21, 2003

Contents

Foreword

This document is written by Mark Burgess, with help and consultation from a variety of sources. It was written in connection with our diversification into higher degrees, as a way of demonstrating the worthiness of the institution to regulating bodies, as well as for our own benefit. The author has been experimenting with these issues for several years an it is clear that the need for procedural security is much greater in the Bachelor degree, where large classes predominate. However, this also affects our ability to deploy higher degrees, so the whole matter of security and quality control is central to all parts of our insitution.

In writing this document I have had close contact with Mari Mehlen, in the Course Quality Committee. She has provided data and comments that have made it possible to be more scientific about the process. "Too many handwaving opinions, too few facts" — this was my attitude in starting this work. The results show that most of the handwaving opinions of both staff and students are incorrect.

Peer reviewers: Mari Mehlen, Jan Kleppe, Cecilie Rolstad, Tore Hoel. *MSc Quality Assurance Committee*: Alva Couch (Tufts University, Boston), Curt Freeland (Notr Dame, Illinois), Morris Sloman (Imperial College London)

–MB

Chapter 1

Introduction and goals

This document is written in English, that it can be examined and evaluated by international peers of Oslo University College for the sake of quality assurance.

Definition 1 *The purpose of this document is to:*

1. *Define and evaluate methods of measuring the quality of student learning.*
2. *Define and compare methods for using grades.*
   (a) *Grades as a motivation / reward for work*
   (b) *Grades as a measure of student achievement*
   (c) *Grades as a measure of student competence*
3. *Quantify the uncertainty (probable error) of the measures above.*
4. *Recommend quality control guidelines, that minimize the uncertainty and maximize student learning, in the form of ISO 9000-style checklists for the various aspects of teaching and examination.*

The principal audience for this document is the teachers and staff engaged in the education of students. The author wishes to emphasize that this is not a political manifesto, nor a contract between any parties. It plays a purely advisory role. This quality of this document is attested to by the signatures of at least four individuals of differing initial backgrounds and opinions.

The goal of the document is to determine ways to increase the quality of learning and evaluate the probable uncertainty in the measurement of student capability. The document is therefore to be understood as being in the best interests of

- Students, to whom Oslo University College owes a quality education.
- Society, to whom Oslo University College owes quality students.

Different methods of teaching and student assessment are often discussed, including:

- Final supervised examination
- Continuous assessment (un-supervised)
- Online testing (multiple choice, peer review)

This document offers guidelines for these which maximize the probability of student learning, and which minimize the error and uncertainty in setting student grades. For the purpose of this document, all grades are measured on a percentage scale.

1.1 Some definitions

Definition 2 *The term "learning" refers to the acquisition of knowledge or skills, of any kind, as a result of time spent at Oslo University College.*

Definition 3 *The term "subject" refers to a learning discipline, as represented by a lecture course or laboratory course etc.*

Definition 4 *The term "degree award" refers to the final average grade of all the subjects required to be taken by a student in order to complete their studies.*

Definition 5 *The word "assessment" or "measurement" is used to mean a method of grading student work. The method of grading is to be specified; e.g. examination, oral presentation, project work etc.*

Definition 6 *The word "examination" is used to mean a written test carried out under supervised conditions.*

Definition 7 *The terms "oral-examination"*[a] *is used to mean a private interview with a student, under supervised conditions, with at least two authorized examiners present, one of which is the course teacher.*

[a] In some countries, this is referred to as a "viva voce" (abbreviated to "viva").

Definition 8 *The term "oral-presentation" is used to mean a lecture or presentation carried out by the student, either publically or privately, which is judged by a panel of at least two authorized members, one of which is the course teacher.*

Definition 9 *The term "group work" is used to mean work carried out by a number of students in concert, where no supervision is maintained concerning which students are responsible for which contribution.*

Basic fact 1 *The grade assigned to a student at the end of a course is a matter of policy. There are relatively few subjects where an exact numerical grade can be used to represent a precise right-wrong distinction. A grade is a signal to the student and to society, but it is also the currency of reward to students, since a high grade confers benefits in later life. These facts must be borne in mind when setting a policy for grading work.*

Basic fact 2 *Most institutions do not explain their grading schemes; grades are taken on* trust. *A quality assurance plan aims to make the foundation of that trust plain to all parties.*

The results of this document are an effort to specify quality methods for serving the needs of students and society.

1.2 Goals

The goal of a University or College is to educate students effectively and efficiently. It is a balance between good economics and optimum effort. Assessment of students has two roles:

- It measures the achievement level of a student.

- It forms feedback loop which informs and motivates students in their progress.

This document must serve both of these.

Goal 1 *To estimate the uncertainty of different assessment schemes so that they can be compared for reliability.*

Goal 2 *To estimate the probable gain in short and long-term competence by various approaches.*

Goal 3 *To estimate the reliability of the grade as a measure of competence, using approaches 1 and 2.*

Goal 4 *To establish uncomplicated procedural guidelines for minimizing the error in assessment methods, due to random and intentional causes, including exploitation of loopholes and 'cheating' by staff and students.*

Certain schemes, such as IQ tests, suppose to measure an ability that does not relate to a particular subject matter. We do not consider these. Grades can be used in two independent ways:

1. As a certification of competence (ability to reproduce knowledge or skills).
2. As an incentive to learn (reward for effort).

As long as reward for student work results in improvement of student competence, there is no contradiction between these uses.

1.3 Standardization

Standardization promotes certainty. Methods of standardization include:

- Procedural checklists that ensure an identical treatment of students within a course.
- Assurances that teachers follow some standard procedure for testing.

Standardization does *not* imply

- Fixing the values of arbitrary parameters e.g.
  - Method of assessment for a given course.
  - Frequency, exact number, weight or length of measurements/tests.
  - Size of classes.
- Assumption that all students are of equal ability.

We do not expect to be able to find numerical data to support every idea in this document. In cases where we are not able to make value judgements, we report:

- Reasons to support based on experience of individuals.
- Reasons to oppose based on experience of individuals.

The document discounts any opinions that are not justified on the basis of experience. One does not wish to engage in speculation that might restrict the freedom of teachers to experiment with innovative teaching or assessment methods.

1.4 The end of exams as we know them?

There is an urgent need to review examination security procedures. The profileration of microelectronic communication devices is quickly making traditional exam supervision methods invalid. It is now almost trivial to secure an undetectable commnications link to almost anyone. In a few years, it will be trivial to cheat in exams using modern telecommunications equipment, *unless an information security strategy is worked out*.

Projects and essays can be bought on the network, or be passed down from year to year. Unless *variations* or *traps* are included in project work, it is not possible to assume that submitted work is actually *fresh*, i.e. not a replay attack. The work of the authors

Unless real security measures are introduced, in an intelligent manner, the grades claimed by Universities and Colleges will no longer be trustworthy to society. This problem needs to be addressed by security experts – not by political dictate.

Chapter 2

Learning

The purpose of this chapter is to highlight some of the factors that are important to student learning.

2.1 Definition of 'Quality of Learning' (QoL)

Maximization of student knowledge is a problem for economics or strategy. We are looking for strategies which strike a balance between maximizing results and minimizing cost. The cost of providing a quality edutcation results from

- A finanical cost in terms of resources per student.
- An administrative burden on teachers and staff.
- The time spent in teaching and supervising.

The goal of quality assurance is to maximize the sum of knowledge over time. Let us assume that the knowledge of a student s can be measured numerically, in a given subject, and be represented by a function of time $K_s(t)$.

Definition 10 (Learning estimate) *Learning quality per student Q_s is cumulative, proportional to time interval $[a, b]$ spent working:*

$$Q_s = \int_a^b K_s(t)\, dt. \tag{2.1}$$

The function $K_s(t)$ is unknown, but it can be assumed to be non-negative.

Quality is not extensive, i.e. Q_s and $K_s(t)$, are not proportional to the numbers of students, thus we do not consider 'total throughput' to be a substitute for knowledge per student in our definition of quality.

Principle 1 (Maximize contact with studies) *To maximize quality of learning, we thus have to maximize the domain and range of $K_s(t)$, i.e. the amount learned and the time for which the knowledge remains with the student.*

What kind of value does $K_s(t)$ represent? We have to be aware that it is not the number of exams passed. Focusing on student examination 'passes' is not an adequate representation of society's need for knowledge, since one can easily find strategies for passing students without student competence increasing.

Knowledge and ability change with time, so we might also consider the efficacy of different strategies for maintaining knowledge with time also – but this is secondary.

2.2 Factors affecting quality of learning K_s

Student attention span (work effort) during a course is widely believed to have the empirical form of the curve shown in fig. 2.1. This hypothesis seems to work as an approximation to the truth both in working towards final examinations and during continual assessment.

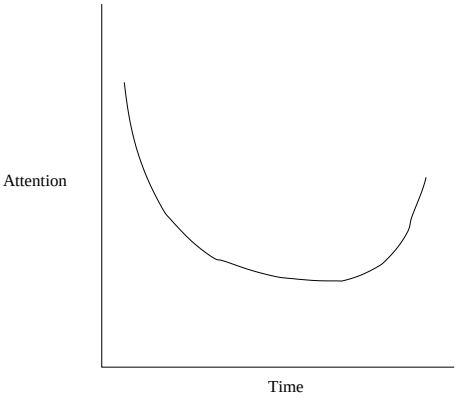

Figure 2.1: Student attention hypothesis over any given interval of time, e.g from the start to end of a lecture, and from the start to the end of the course. Experience bears out the shape of this curve, though I do not know of any data measuring this...

We can use this curve as a basis for time-planning courses and lectures.

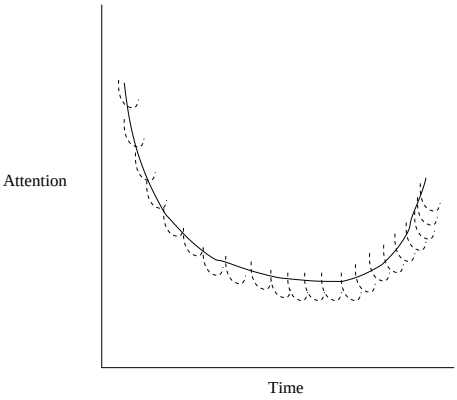

Figure 2.2: Detailed view of learning hypothesis over the course teaching period.

Students can reduce the quality of their learning by cheating. Cheating is a strategy for misrepresenting learning level, for the gain of another type of social currency, such as recognition or peer status. We do not discuss the economics of cheating, i.e. what students really gain by it, except insofar as to consider ways to make it worthless for students to cheat, i.e. to reduce their perceived gain.

The language of games is useful. We define some terms.

Definition 11 *The term* attack *is used in the sense of gaming. An attack on a system is an exploitation of a weakness in the system.*

We use this word to generically describe all failures of system procedure and security, no matter how they arise.

- Most students agree that the amount of learning is increased by work, by time spent on work (since the effect is cumulative), and by effort expended.

- Effort invested by students is increased by the expectation of difficulty, provided there is a chance to succeed; thus study tasks should not be made too easy.
- Learning effect is reduced by distraction, loss of interest, illness, loss of respect for the course.
- If the social reward to cheat is greater than that to succeed legitimately, it is worth the student's while.

Principle 2 (Teaching strategy) *A strategy for maximizing learning is to raise expectations, i.e. increase the pressure, raise the standard, while providing a way to succeed.*

2.2.1 Student ability

We assume that it is possible to measure and represent student ability. No measurement scheme can be made 100% free of uncertainty.

1. Interest level, motivation.
2. Work discipline
3. Natural ability.

2.2.2 Teacher ability

Teacher ability is presently a significant factor in student learning. One of the aims of this document is to suggest routines for learning which reduce the dependence on the teacher. In a sufficiently assured scheme, students would not be completely dependant on a given teacher. This would provide security against.

- Teacher absence
- Teacher incompetence
- Personality conflicts

2.2.3 Change of student ability with time

Student ability is not constant. It starts from somewhere near zero and grows with experience. In periods of inactivity, it can also decline again. Ability falls into two types:

- Instinct: a trained response which does not require reasoned thought, e.g. playing musical instruments, driving a car, simple differentiation, certain programming skills etc. ("It's in the fingers") This type of ability tends to persist for a long time, and is based on frequent repetition.
- Cognitive: a skill which requires a reasoned consideration. This is more fragile knowledge that disappears more quickly. It is not based on repetition, but on analysis.

One has two strategies for maintaining ability:

- Repetition of frequently used skills.
- Practice in analysing new variations.

Chapter 3

Student assessment methods

The purpose of this chapter is to examine data about the reliability of different measures of student achievement. There is a broad range of ways in which student performance can be measured.

3.1 The reliability of an assessment method

The method of probabilistic error analysis can be used to calculate the uncertainty in grades set by different measurement schemes. There are two issues

- The reliability of the procedure to accidental error.
- The security of the procedure to attack.

We require some basic estimates of the likelihood of certain events taking place (e.g. cheating). These can be obtained from data available to us from previous experience.

A cause-effect tree, or fault tree, has the form of fig. 3.1. We can create such trees of cause and effect to analyse the probability of achieving the goal of quality. We shall

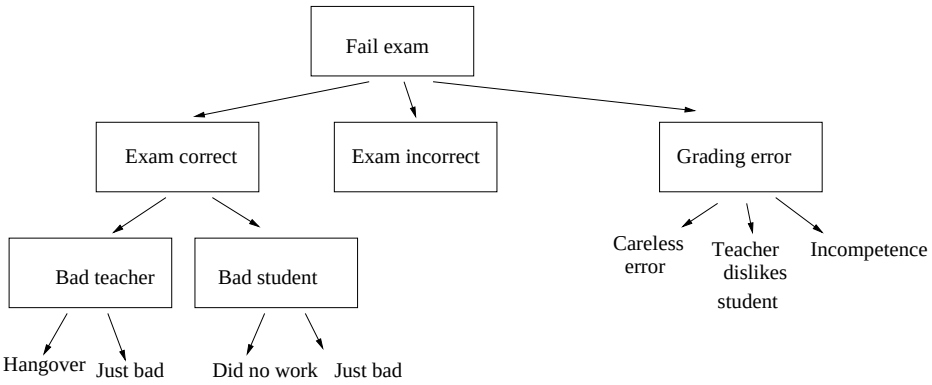

Figure 3.1: A partial cause tree example for examination failure. Many more factors can be listed, but what are the most important ones?

consider two aspects of quality:

- Quality of result achieved (student grade performance).
- Quality of examination and grading procedure (teacher marking performance).

Note: we disregard assumptions about the distribution of student abilities within a class, e.g. such as normally distributed abilities etc, since these defeat the idea that quality assurance can improve student learning. Such assumptions play no role in this analysis. See however section C.2 for comments on this.

3.2 Scalability cost of an assessment method

Quality control clearly has economic limitations. We look therefore for a compromise, and adopt the following principle in recommendations. If a little money can lead to a large improvement, it is worth the cost. If a lot of money leads only to a small improvement, it is not worth the cost. Anything in between is a matter for policy.

For example, if one has only a handful of students, it is possible to interview each one of them before a broad committee. This offers a high level of impartiality and security. The uncertainty in the result is then a result of student disposition (nervousness, health and environmental concerns etc). This method is expensive in both time and resources however, and does not scale to hundreds of students.

Parallelization is a strategy that scales, while *serialization* is a strategy that prevents scaling. Thus if testing can be performed in parallel, it will scale freely:

- Any electronic exam scales to arbitrary numbers of students.
- Peer review scales to any number of students.
- Paper exam with multiple graders can scale if there are sufficient graders (but leads to a high level of uncertainty in results due to the differences between graders, and is expensive).

Serialization (a bottleneck) does not scale:

- Graded exam with internal and external examiner channels all work through one or two individuals.
- Oral examination must be carried out one at a time.

3.3 Attacks on assessment methods

A certain percentage of students must be considered hostile. That is, a certain number of students will attempt to exploit loopholes, cheat and otherwise gain a grade without playing by the rules.

One can try to remove their incentive to do this, for example by using a pass/fail grade, or by otherwise indicating that grade does not matter, but this also removes their incentive to succeed. There is thus an inevitable risk associated with grading.

To ensure that all students are treated fairly, assessment questions must be secret in advance of the test and test papers must not be available to students after the test; i.e we must ensure

- The security of questions prior to test.
- The security of answers after the test.

During a test students can cheat in three ways:

- By copying from books and other information sources (factual).
- By getting another person to answer the question for them (authenticity).

- By replaying the responses of another who has already taken a similar test (replay attack).

The elimination of covert channels of communication between students and others during a test measurement is usually important, but other strategies can make it difficult for students to communicate, e.g. pressure of time to complete a test, randomization of questions etc.

Electronic methods can help to eliminate random or human error, by forcing a more rigid framework upon examination, and by making students responsible for their own registration (parallelism). Any serialization of data-flow (through a single individual) increases the risk of error. However, even with electronic methods of registration there can be serious errors. Unreliability of data transfer, e.g. form bugs in software (e.g Internet Explorer) can lead to incorrect results being registered; moreover, regardless of how reliable a computer might be in grading flawlessly, the *human management* of a test is still paramount in controlling the human environment.

From informal interviews and investigations as part of our course on security, we can estimate that, in recent examinations around 5% of students cheated. In a recent electronic examination, a similar number of students were able to cheat when the physical test environment was spoiled by server failure. Clearly not all problems can be foreseen, and one needs a *contingency plan*.

Some problems in human management include:

- Exam monitors are often pensioners who are naive about students and security, and have no authority or guidance about security practice.
- Toilet breaks are frequently used to pass information in and out of exams.
- Students can collaborate in duping the examination security, and some students even have developed well-oiled machinery to do this.

The main conclusion one can draw from this is that human management is the main cause of security breaches, and available data show that neither written examination nor electronic testing is more secure.

3.4 Reduction of procedural errors in student assessment

1. **Peer review:** used for human work (both in checking assessment design and assessment results). This helps to eliminate bias, random error, systematic errors and opinion bias in individuals. It is not immune to common misconceptions amongst the peers.
2. **Machine checking:** used for electronically submitted tests.

In all cases where students are graded *èn masse* (but especially where they grade each others' work) one should include control questions and dummy answers in the mass that should fail to obtain an acceptable grade. Similarly, but less useful, one could include an ideal answer. These points of reference allow grade scales to be scaled and adapted to the uncertainties inherent in the process.

3.5 Timing of tests

It one attempts to measure ability too late, after completing a course, the result will show signs of forgetfulness, as they put aside old material which they do not currently give priority. This can be minimized by averaging over time, i.e. continuous assessment. One would aim to see a logarithmic progression in students' achievements over time.

If one attempts to measure ability too early, slower students will not have had time to mature in their understanding of the material and an artificially low grade will be registered.

3.6 Uncertainties in assessment methods

In this section we attempt to evaluate approximate numerical values for the probabilities for system failure, using the various testing methods.

3.6.1 Uncertainty in multiple choice assessment

The uncertainty in a multiple choice exam is very straightforward to address. It comes from two factors:

- Monkey mark (success by random guessing)
- Question ambiguity

In a multiple choice examination, any student can achieve the 'monkey mark' by guessing at random, unless policy stipulates otherwise.

1. **Multiple choice with no penalties:** If the average number of choices per question is $\langle N \rangle$, then the monkey mark is $100/\langle N \rangle\%$. Accordingly, the uncertainty in the multiple choice procedure is the same.

   The probability of success for a student becomes the probability of an error in the examination integrity for the assessment:

   $$P(MP) = \frac{1}{\langle N \rangle} \tag{3.1}$$

   This can be reduced by having a larger number of choices per question. The uncertainty accorded by question ambiguity is of the same order of magnitude as the monkey mark. The number of answers that might be chosen by a misunderstanding is a number between 1 and $N - 1$ (hopefully, at the lower end rather than the upper end). With a proper quality control of the questions prior to the test, we can assume that the probability of choosing incorrectly will be minimized, i.e. $1/N$. The combined uncertainty is thus

   $$U = \sqrt{(1/N)^2 + (1/N)^2} = \sqrt{2}/N. \tag{3.2}$$

   Thus, if N_i is the number of alternatives in question i, and w_i is the weight accorded to that question, scores on multiple choice tests should be interpreted as

   $$\text{Score} = \sum_{i=\text{correct}} w_i \pm \sqrt{\sum_{i=\text{all}} \left(\frac{\sqrt{2} w_i}{N_i} \right)^2} \tag{3.3}$$

2. **Multiple choice with penalty for wrong answers (non-negative):**

   If students receive a negative grade for giving an incorrect answer, the effective monkey-mark is reduced to zero. Now the uncertainty of choice attacks the student's security however, since there is a greater number of incorrect answers than correct answers. The chance of making a mistake is is the same, but the penalty is not in the student's favour.

3. **Multiple choice with penalty for wrong answers (allow negative):**

   This is the same as case above, but the total test score can be negative in total, so that the test actually subtracts from the total of all tests. This could be used to compensate for too high a grade in earlier tests that are non-supervised (e.g. group work).

   Example: project work that is performed in groups can be combined with a test of this kind that probes the nature of the project, so that those who did not participate or contribute to the work are penalized for their ignorance of the work they claim to have submitted.

Note that poorly crafted multiple choice questions can be guessed without requiring any knowledge of the course materials. 'Trivial Pursuit' questions such as 'what does TCP stand for?' have little pedagogical value and can be guessed quite easily on the basis of likelihood.

3.6.2 Uncertainty in human graded assessment

In an examination graded by humans, there are many sources of error. Here are some examples:

- Careless error (random error)
- Difference of opinion
  - Grader mixed up papers
  - Grader works too quickly
  - Grader fatigue
  - Grader interruption/confusion
- Dislikes student (attack against student)
  - Personal conflict
  - Prejudice issue (race,sex)
  - Failed previously
- Grader incompetence
  - No knowledge of course
  - Improperly read question
  - Two examiners agree on wrong ans.
- Procedural error
  - Papers mixed up
  - Computer data error
- Unreadable handwriting
- Exam was too easy
- Student cheated (attack against College)
  - False identity
  - Copying during exam
  - Covert information channel
  - Ghost-written project work

The probabilities of some of these occurrences are difficult to estimate. There are two main types of assessment to distinguish:

- **Unambiguous answer assessment**: this is generally easier to grade. Here is it found that this is of the order of 5% (see appendix C). This error cannot be estimated, it is given on the basis of the best available empirical data.

$$p \sim 5\% \pm 1\%. \tag{3.4}$$

- **Value judgement assessment**, essay, oral examination or otherwise ambiguous answer: this involves a value judgement by an examiner, and is thus open to wider uncertainty.

  1. In project work where examiners have been asked to grade exams on general criteria (e.g. hovedprosjekt, system administration projects), with no strict checklist, the grade deviation can be quite large.

$$\langle p_1 \rangle = \langle \sigma \rangle \pm \max \sigma \tag{3.5}$$
$$\langle p_1 \rangle = 8\% \pm 24\% \tag{3.6}$$

     (Data from computer science final projects 2001 and Network and System Administration projects over several years.)

  2. When project work is graded according to a checklist, the deviations of grades in project work can be very large when reviews are set by a single inidvidual. Estimates from project work, with reviews taken in threes, give:

$$\langle p_c \rangle = \langle \delta \rangle \pm \max \delta \tag{3.7}$$
$$\langle p_c \rangle = 5\% \pm 18\% \tag{3.8}$$

     In other words, the expected error is about 5% by a competent grader, but there is a potential error of up to 18% in the worst case.

  3. When checklisted projects are graded in threes, the uncertainty falls by virtue of greater concensus:

$$\langle p_3 \rangle = \langle \sigma \rangle \pm \max\langle \sigma \rangle \tag{3.9}$$
$$\langle p_3 \rangle = 1.5\% \pm 9\%. \tag{3.10}$$

     (Data from 150 students in computer security, rounded to nearest whole number)

There are several things we can do to minimize the errors in these examinations:

1. **Aim for a concensus of several opinions:** This is problematic in itself if the opinions are not homogenized in their knowledge of the course details. An external examiner with no knowledge of the course contributes to a grade essentially as a random error.

   In a peer review scheme, where students mark each others work, one is fairly certain of what students know about the course, plus or minus a random scatter. One can reasonably assume that the scatter is symmetrically distributed about some average value so that a reasonable number of opinions from a peer review group will not be unfairly biased, and will have an uncertainty characteristic of the normal scatter in the class.

   There is no evidence to suppose that work graded by a teacher in more reliable than work graded by an average student. In large classes, teacher fatigue is a problem, and multiple teachers leads to new uncertainty.

2. **Anonymity – double blind marking:** Anonymity helps to remove malicious attacks by staff against students, and malicious complaints against staff by students[1].

   In a peer-review situation, anonymity helps to protect students from malicious attacks. However, it cannot really be achieved in groups of less than 50. Students have a sufficient number of social ties as to be able to know or guess whose work they

[1] NOTE: MS Word documents should not be used for submission, since they contain hidden references to student names. PDF fi les should be used for all written work.

are grading. Some students are clearly identifiable by their ability, language idiosyncrasies or other factors. Thus fairness cannot be guaranteed. Handwritten tests easier to identify - teachers recall maybe 5 students from a class?

Recommendation 1 *In project submissions requiring anonymity, MS Word documents should not be used, since they embed the author's name wihtin the document, providing a covert channel to the identity of the author. It is recommended that documents be converted to PDF format in all cases.*

- Openness of procedure provides extra checks: Students can check their own results. Perhaps 90% of students accept an explanation of the grade they received. Most complaints arise from a lack of understanding of their grade.
- Fast feedback is important: A slow reply aggravates feelings of complaint.

3.6.3 Oral examination

Although oral examination is a human graded assessment and the above applies, oral examinations warrant a special mention, because they are often performed on a one-off basis. Grades are often set by 'feeling', as a matter of policy. The uncertainty is very high, and is usually at the whim of the course instructor. We have no specific data on this at the College since this type of examination has been rare.

Oral examinations should be conducted by a strict checklist as far as possible, in order to minimize the level of uncertainty.

3.6.4 Security of group project work

There is a large uncertainty in grading project work. Project work remains popular amongst teachers however, for its pedagogcal benefits. Students who carry out project work in the intended manner invest more effort in and learn more from this type of 'test'. The uncertainty in grading project work is the same as that in the previous sections, with an additional uncertainty for participation by all members.

The uncertainties are:

- Knowing who is responsible for the work: students could hire a ghost-writer for the project. This is difficult to detect.
- In project work, environmental controls are rarely possible.
- In groups, not all of the students need not work in order to get a grade ('free riders'). This can be countered with an individual test about the project work itself.
- Project work is often of a subjective nature. Grading error depends upon the opinions of the examiner. Fairness thus requires an averaging of opinions from several examiners. Peer review is suitable for this task, or a committee of examiners.

Countermeasures against project cheating:

1. **Personal supervision** during the project allows a supervisor to gain an impression of who contributes to the work; however, this is expensive in teacher time.
2. **Oral examination** or 'viva' of each student is the most revealing form of test; however some students give a poor impression through nervousness. Also this is a very expensive method of testing that only works for small numbers (less than ten students per day, over a small number of days) before staff become exhausted.

3. **Crafted (online) test** that asks questions about submitted work. Time limitation combined with simultaneous randomized problems can replace test supervision. (See procedures, below.)

4. **Voting:** students vote on each other's participation. This is subject to "mobbing" attacks by students who 'gang up' on a minority, or conversely students protect one another; however, such cases are rare and could be dealt with by a staff referee (adjudicator).

Project work can be graded by peer review, or by a teacher who can interview the students to make sure they know and understand the content of their project. Who actually wrote the project is not of any interest to this result: we are simply interested in whether the students have understood what they claim to stand for.

Project grade uncertainty per student consists of two parts: a general part due to grading error and a part due to likelihood of grade over-estimation as a result of free-riding by group members.

$$p_{\mathrm{proj}} = p_{\mathrm{grade}} + \frac{(N-1)}{N_{\mathrm{estim}}}\, p_{\mathrm{free}}. \tag{3.11}$$

Here N_{estim} is an estimate of how many persons are required to carry out the work, with a reasonable workload, and N is the actual number of students. This expression is an expression of *policy*, based on *speculations* about student ability and integrity. It can be used as a model and supported by empirical data in the future.

As an order of magnitude estimate, perhaps 10% of students allow others to work for them at some time.

$$\langle p_{\mathrm{free}} \rangle = 0.1 \tag{3.12}$$

Since the uncertainty due to grading alone is of the order of 5-10%, the uncertainty due to group work can play a significant role and even dominate the grade.

Many students work in groups and share workload. This is not a problem unless it reduces learning significantly. How seriously one regards the problem fo project uncertainty is a matter for policy.

Recommendation 2 (Guidelines for grading) *Clear guidelines for marking are the most effective way to ensure grade uniformity against human error.*

Project work grades should be set strictly by peer reviewed checklist, and should include countermeasures against free-riding, or the grade should be pass/fail. Students can be given a verbal indication of their performance.

The alternative to the above is a high degree of uncertainty.

3.6.5 Uncertainty in electronic exam assessment

The uncertainties in electronic assessment come from:

- Human error in questions.
- Uncertainty in the identity of the student being tested.
- Bugs in the submission system.
- Security of the questions.

In addition to this there is the need to know the true identity of the person taking the test, and the fact that the work is their own. These two issues require *human supervision*. Even with the various logs and electronic watchdog mechanisms available, one cannot be certain of identity unless a physical meeting takes place. Such a physical meeting can be used to issue a *one-time password* for an exam. Regular daily passwords cannot be trusted.

3.7 Combining uncertainties in the final grade

The uncertainty in the assessment methods above can be calculated as the probability of failure multiplied by 100 (percent). This represents the resolution with which it is meaningful to distinguish between grades.

The estimates of the probabilities P_i themselves could be in error, by an amount $P_i \pm \Delta P_p$. These errors also contribute to uncertainty, which can be calculated by the standard theory of errors. Given that the probabilities of different sources of error are independent, their contributions may be added or multiplied, by well known rules;

$$P(A \text{ AND } B) = P_A P_B \tag{3.13}$$

$$P(A \text{ OR } B) = P_A + P_B - P_A P_B \tag{3.14}$$

The uncertainty of a sum or a difference, for independent contributions is calculated by the Pythagoras rule:

$$\text{Error}(P_A \pm P_B) = \Delta P = \sqrt{(\Delta P_A)^2 + (\Delta P_B)^2 + \ldots}. \tag{3.15}$$

The uncertainty in a product $P_A P_B$ is

$$\text{Error}(P_A P_B) = \Delta P = \sqrt{P_A^2\,(\Delta P_A)^2 + P_B^2\,(\Delta P_B)^2 + \ldots}. \tag{3.16}$$

The probability of error is thus $P \pm \Delta P$, thus the uncertainty in the grade can be taken to be:

$$\Delta G = 100\,(P + \Delta P), \tag{3.17}$$

and any grade should be specified to students as an actual assessment, plus or minus the uncertainty, with the average grade quoted as a reference.

$$\text{Grade} = G \pm \delta G \tag{3.18}$$

The grades should not be classified in intervals of less than δG.

3.8 Ranked list of uncertainties

In order of decreasing certainty:

1. Definite answer electronic grading
2. Multiple choice electronic grading
3. Peer review with at least three graders
4. Internal plus external examiner with unambiguious answers
5. Single grader
6. Group project work

Oral examination cannot be placed in this list. If not checklisted, it is pure value judgement uncertainty, and oral exam grades are usually a matter of policy or 'feeling'.

It is well known from game theory that a mixed strategy is often the best approach. In this case a mixture of rewarding effort and punishing incomprehension provides both a fair scheme to students and a reliable one for the College.

Recommendation 3 (Mixed testing strategy throughout courses) *Courses should begin using the strategy of grade for reward, but make use of a strict test that is designed to reveal 'free-riders' and compensate for false credit acccumulation.*

The security of the final test is paramount in moderating grades and achieving the College's aims of grade reliability. Even electronic tests require human supervision. The test should be in the latter half of the course, but not necessarily at the end. Students will be motivated to compensate for their poor test performance.

Chapter 4

Reward methods for learning

If a University College is to maximize its goals, then it must maximize genuine student learning. Testing students does not increase their ability, it only provides a weak incentive to learn. There is thus a trade off

- A college should choose strategies that maximize student learning, in order to fulfill its obligations to the students and to society.
- A college should measure student ability as a quality control on the learning procedure, and as an approximate signal of achievement to society.

4.1 Continuous assessment vs final exam

There is real evidence from repeated student questionnaires to support the idea that continuous assessment leads to increased effort by students. Continuous assessment shifts the time at which student begin to work to an earlier stage, thus extending the time interval over which they are actively involved in studies.

In addition to actual error in grading, there is an intrinsic error due to normal social development of students. Every student matures with exposure and practice. If one attempts to measure their ability too soon, the result can be too low. It one attempts to measure ability too late, after completing a course, the result will show signs of forgetfulness, as they put aside old material which they do not currently give priority.

4.2 Example weighting scheme for continuous assessment

A natural way of weighting measurements of student work, when assigning grades is based on the idea that ability and understanding increase approximately exponentially with time, i.e. the propensity for new learning is proportional to what they already know K. This is simplistic, but can be used as a rough guide. Suppose,

$$\frac{dK}{dt} \propto K, \tag{4.1}$$

so that

$$K(t) \sim K(0)e^t. \tag{4.2}$$

A logarithmic weighting of scale for grading, of the form,

$$W = W_0 \log(1 + t/T) \tag{4.3}$$

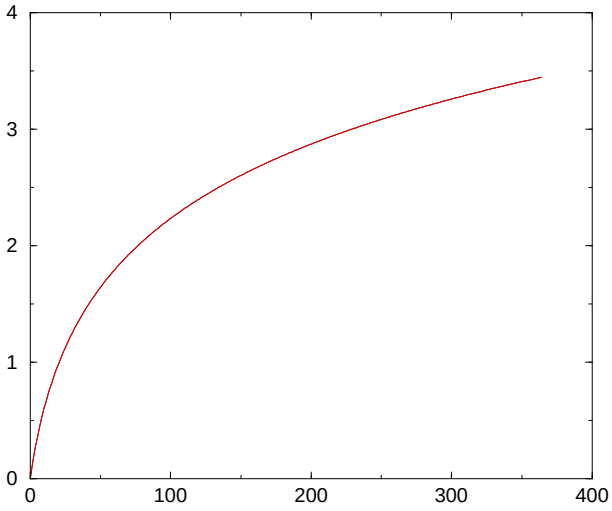

Figure 4.1: Plot of W/W_0, over the course of a year, showing a logarithmic form of grading. If knowledge is compounded, then the rate of growth is approximately exponential. This curve is a simple guide for weighting the quality of student work at different times throughout a course.

is this a simple weighting policy which reflects the compounding of knowledge and experience, of increasing importance (see fig 4.1). Less weight is given to what an inexperienced student achieves, and more is given as experience is gained.

The curve in fig. 4.1 is somewhat naive, and probably idealized. Since attention span and work effort follow the form of fig. 2.1, the curve is likely much flatter than this, perhaps even linear.

4.3 Ability after learning process

Knowledge and ability decays with lack of use. We would like to minimze the rate of this decay, if possible. We might assume that the rate of decay of knowledge is inversely proportional to the time $T_{\rm L}$ spent on learning or acquiring it:

$$\frac{dK}{dt} = -K/T_{\rm L}. \tag{4.4}$$

This is clearly a generalization, and it is probably not a useful priority at this stage. Can we choose teaching strategies that minimize this?

- Lecture on situations that students will meet later – this will stimulate memory through repetition.
- Refresher courses.

There is no evidence that either continuous assessment or final examination results in increased effort, but student surveys over several years show that students themselves believe that continuous assessment helps them to remember material better. Based on the model of learning through repetition and contact, this is likely to be true, though no data are available to quantify the assertion.

4.4 Uncertainty in strategies for maximizing learning

We now want to compare the likelihood that one of two strategies

- Final grade used as a reward/incentive for learning
- Final grade used only as a certification of competence

will achieve a high quality of student learning, over long periods of time. Since the currency of work and examinations is grade average, we are really asking two questions:

- What is the probable uncertainty in grading accuracy,
- What is the probability that grade reflects real learning?

For simplicity we can categorize the probabilities as if they were independent (mutually exclusive) events. This is seldom the case in reality, but it is a convenient calculational aid.

Advantage	Exam with ext A	Exam no ext B	Timed online C	Term project D
Student worked hard				
By own choice	$p =$	$p =?$	$p =?$	$p =?$
Forced by college	$p =$	$p =?$	$p =?$	$p =?$
Inspired and motivated	$p =$	$p =?$	$p =?$	$p =?$

Rapid feedback on work quality, and rapid correction. Students should aim for perfection - and we should reward them for correcting mistakes.

4.5 Uncertainty in quality when using grade as a reward

The problem here arises because reward methods favour effort rather than achievement. The contribution to grade is pure uncertainty, but the contribution to quality of learning can be high.

In terms of reliable measurement of ability, the uncertainty is very high and needs to be reduced by combining reward methods with tests that penalize lack of understanding post factum, i.e. tests that reduce the overall grade to redress the balance.

An attack on reward methods can occur if some students are able to improve their grade by talking to students who have already answered tests and finding out what was wrong. It is thus important that, if not all students take the test simultaneously, this form of grade 'attack' should not be able to contribute a significant amount to final grade. It will contribute to learning however, since the students are motivated to find out the correct answer.

4.6 Attacks on reward methods

The main problems with reward based learning is that it artificially inflates grades. There must be a counter-balance in the form of a controlled test that brings the average grade down to managable levels. Significant problems in grade reliability are due to pre-meditated attack.

- Hangers on, or free riders
- Copying and ghost-writing is an attack.

4.7 Conclusion

The main conclusion to be drawn from this is that one should not take grades too seriously. Students evaluations indicate that tghe incentive to work continually is very effective for the majority of students.

It is possible to distinguish achievement within a class, but as a scale of absolute correctness, grades perform quite aritrarily, depending on the assessment methods used.

Principle 3 (Assessment uniformity) *The reliability of an assessment method lies in its fairness in comparing students, not in setting an absolute value to grades. The key point to assure is the predictability of the result, by fixing a well-defined procedure and sticking to it.*

Chapter 5

Quality assurance guidelines

5.1 ISO9000

The ISO 9000 series of standards represent an international consensus on management practices that apply to any process or organization. The aim of the standards is to provide a schematic quality management system and a framework for continual assessment and improvement. ISO 9000 has become quite important in some sectors of industry, in the countries that have adopted it.

First published in 1987, the ISO 9000 standards are widely used and, a quick search of the net reveals that they are also a money-spinner. Courses in these methods are numerous and costly. The principles, however, are straightforward. THe idea is that a standard approach to quality assurance with leads to less uncertainty in the outcome. Quality is associated with certainty. Here, we shall not dwell on the issue of ISO 9000 certification, but rather on the guiding principles that the standard embodies.

ISO 9000 reiterates one of the central messages of system administration and security: namely that these are on-going processes rather than achievable goals.

- *Determine quality goals*
- *Assess the current situtation* (previous knowledge)
- *Devise a strategy* (course plan)
- *Project management* (lectures)
- *Documentation and verification*: (course materials and examination procedure)
- *Fault handling procedure*

How we carry out a process is at least as immportant as the process itself. If the process is faulty, the result will be faulty. Above all, there must be progress. Something has to happen in order for something good to happen. Often, several actors collaborate in the execution of a project. Projects cost resources to execute — how will this be budgeted? Are resources adequate for the goals specified?

One of the key reasons for system failure is that systems are so complex that their users cannot understand them. Humans, moreover, are naturally lazy, and performance with regard to a standard needs to be policed. Documenation can help prevent errors and misunderstandings, while verification procedures are essential for ensuring the conformance of the work to the quality guidelines.

5.2 Checklist for quality assurance in course design

1. Does the course challenge the students to achievable goals?
2. Does the course start at the right level, and make contact with previous courses?
    - If a course in too easy, students will lose respect for the course.
    - If a course is too hard, students will write the course off as pointless.
3. Does the course capture students' attention? Does the course paint a picture, or tell a story?
4. Does the course make contact with current events and situations that students can related to?
5. Have you provided a road map so that students always know where they are going, how far they have come, and why?
6. Does the course push the students constantly towards their goals?

Recommendation 4 (Contract with students) *Each student should be made to sign a contract for each course, indicating that they understand what is expected of them, and accepting the course model. This allows courses to following varying models, and removes the need for an unreasonable uniformity in course design for the sole purpose of avoiding student complaint.*

5.3 Checklist for quality assurance in lectures

1. Is the point of each lecture explained?
2. Does the lecture leave the students wanting to find out more?
3. Are your notes appropriate? (Not too detailed, or too sparse.)
4. Is the use of visual media tidy and well organized?
5. Does the lecture keep students busy and alert or does it make them passive and send them to sleep?

5.4 Checklist for quality assurance in weekly course problems

1. Weekly problems should be related to the lectures of the week and should follow the following template and checklist:
2. Students should meet challenges that they can overcome. They should succeed gradually at tasks they did not know they could succeed at.
3. Each set of problems should begin with an explanation of the purpose of the problem, i.e. what the student can expect to learn by doing them. "The aim of these problems is to...."
4. Gratuitous questions should be avoided, even for practice purposes. They irritate students and lead to complaints and confusion.

5. Each set of problems can contain a list of self-test questions which summarize the main points of the lecture that week. This aids comprehension of the lecture. Hints can be provided.

6. Actual graded problems should show an indication of how much they count towards the final result.

5.5 Checklists for minimizing uncertainty in exams

The examination process contributes a considerable uncertainty to the grade achieved by a student. The uncertainty principle for examinations is that 'we don't know how to rank student achievement without measuring them, but the act of measuring them tends to affect their performance'. Sometimes examinations make students nervous, sometimes they make students cheat. Examination procedure needs to address these issues.

5.5.1 Checklist for exam preparation

The aim of an exam should not be to 'trick' students into answering incorrectly. An examination should present no surprises to students. An examination that cannot be quickly understood by an average student is to be considered of poor quality.

For each question, the teacher should check:

1. Fix a well-defined procedure for grading the course. The procedure should aim to be so well defined that the same result would be obtained regardless of who graded the course.

2. In writing questions, there should be a mixture of three types of question (suggested percentages, based on grading scale, in brackets):

   (a) Problems that can be answered on the basis of attendance to course work (40%).

   (b) Problems that test understanding on various levels (40%).

   (c) Problems that can only be answered by accomplished students (20%).

3. Is there a clear goal to the question? (If not, think of one or delete the question)

4. Can the question be misunderstood? (If so, rewrite it.)

5. Make sure that what is expected of the student is clearly stated.

6. Is the question related to what has been taught in class? If not, make it clear in the question asks something unusual. This kind of question should be limited.

7. Can the answer to the question be graded meaningfully? (If not, rewrite it so that it can.)

8. Write a list of points you are looking for in the answer.

9. Does the question require understanding and thought or just memory? Consider eliminating parts of a question that rely on memory. (e.g. questions that ask users things like 'Which of the following is the temperature at which water freezes'?)

5.5.2 Checklist for exam quality control

Once the exam is written, it should be checked by at least one other person. Peer review of examination questions is important to improve comprehension and eliminate trivial and non-trivial errors. Tests should be checked by an impartial source for ambiguities and other errors. An open source model for course materials can lead to a fast peer review.

5.5.3 Checklist for exam integrity (during the exam)

The human management of tests is essential, regardless of whether the test is electronic in nature or handwritten. Oral examinations are tightly controlled by nature, but mass examination requires concerted effort.

1. Access to the exam must be protected before and after the scheduled examination time.
2. Students must not be able to see or alter examination content prior to the test, and must not be able to change the exam or their answers after the examination date.
3. A list of *critical dependencies* should be determined in relation to the test: how can the test fail to perform its function? Measures should then be determined to avoid these: e.g. if an electronic exam is being used, server performance might be critical, timing might be critical. If the necessary performance cannot be guaranteed, the redundant solutions need to be found.
4. Toilet breaks should be avoided. Short examinations are preferable.

5.5.4 Checklist for grading tests

The key to consistent grading is in establishing clear guidelines for what is acceptable and unacceptable. Weighting systems for different parts of a test must be applied consistently by every individual or system grading the assessment.

Guidelines should be formalized and even automated if possible. This should have been worked out in the course preparation. A computer based registration form for grades is desirable, since this forces every examiner to use the same standard. If the detailed grades are available to students, the probability of queries and complaints will be reduced.

1. Follow guidelines for marking exactly.
2. Give a strict score for each point.
3. Allow bonus points for student remarks you have not anticipated.

Recommendation 5 (Definite answer grading) *Questions with a right and a wrong answer should be answered electronically to eliminate human error. If a problem has a correct answer, then there is no reason not to automate its grading.*

Recommendation 6 (Eliminiation of bias) *If a question does not have a simple, correct answer, one must seek a mixture of qualified opinions. It is economically impractical to achieve statistically significant mixture of opinions, so the opinions should do their best to eliminate bias.*

Recommendation 7 (Verification of procedure) *Include a number of dummy submissions in the mass of items to be graded that should fail to pass the test, or pass with only a weak grade. If these tests do not fail then the grading scheme is incorrect and required tuning.*

Recommendation 8 (Benefit of the doubt) *If a student is in the grey area between two grades e.g. A and B, and the percentile distance to the grade is less than the uncertainty in the exam, the higher grade should be awarded to the student.*

5.5.5 Personal supervision of tests

Personal supervision has an important psychological value to students. While this form of supervision has worked for many years, the communications revolution will quickly invalidate of this form of exam. The difficulty of securing a long examination, using only human supervision is increasing rapidly. The longer a student has to cheat, the greater the chance of their succeeding.

Any devices should be brought to an examination must be regarded as untrusted; they can either be excluded (which is expensive to achieve) or factored into the uncertainty equation. Any tools required should be provided by the examination itself.

One should be aware that, as technology races forward, devices resembling hearing aids and other audio devices could be used to send and receive message. Students who use hearing aids should remove them for the duration of a written examination.

Oral examinations should be carried out in screened rooms, if there is a suspicion of cheating. Small audio receivers are now easily made.

5.5.6 Electronic supervision of tests

Electronic testing has many attractive features: it is easily graded, it scales to large numbers of students and it requires little or no administration. For most students, electronic testing offers sufficient security; however for a persistent and dishonest attacker, it presents an easy opportunity.

The main problem with electronic testing is the culture of ‘carelessness’ that students have learning in connection with computers. They are used to the idea that they will be protected from harm no matter what they do. This leads them to take choice-based tests unseriously – and they make many mistakes.

Current Web technology is poorly suited to this kind of security because it has no personalized or persistent sessions. Even if a web test is randomly generated, students can save a sessionless page for later, thus freezing the randomness. They can then submit the same test for all students with identical responses. There is no way to detect this kind of attack, if students are persistent.

1. Electronic assessment can be timed and randomized to minimize the possible gain by any students helping one another during a test. If there is a short time limit, then a student that helps another stands to lose out personally since the extra work will simply waste their time.

2. One-time passwords or challenges can help to defeat password sharing attacks, but the problem is that all such methods assume the integrity of private identity. One time passwords should be provided at the entrance to the supervised area.

3. Secret identity keys could be handed to students in person, immediately before taking a test, provided they are not given an opportunity to hand over this secret to an imposter. This requires a physical administration for maximum security (see below). Keys could be E-mailed, but this offers a low level of security.

4. If randomized tests are used, each test should expire after a short time to prevent students from replaying earlier test results.

5. The time of examination submission must be recorded by reliable time source.

The central problem in online examination is not covert channels of communication, but rather covert swapping of identity.

If one assumes that students copy mainly from students they choose to work with in project groups, then this information can be used at examination time to minimize the likelihood of a student copying during the test, even without external supervision.

Principle 4 (Isolate peers) *It is not essential to supervise all students, but a random sample should be supervised, plus any students suspected of being 'free riders'. This will tend to break up groups of peer associations.*

Students who work share a peer association, e.g. those who work in the same project group, are the most likely candidates to help one another and therefore the priority should be to prevent these groups from communicating. Some possibilities:

1. *The group should be kept apart, with examination conditions that ensure that communication is not in their best interests - e.g. short time limit.*
2. *All peers can be required to take a test simultaneously with a short time limit. Note that not all groups need take the test precisely at the same time. The greatest risk comes from peer associations.*
3. *Students that finish the test should be 'removed from play' by asking them to report in to a place removed from the examination area. This will minimize the risk of students communicating with new students who have not yet taken the test.*

The security of this type of test descreases as the time limit for the test increases. Unsupervised tests need to be relatively short in order to maintain security.

Recommendation 9 (2 testing strategies) *In electronic tests used for reward credit, no supervision is necessary. Students should rather be encouraged to work together and help one another.*

In electronic tests used as a punishment strategy, as many students as possible should be supervised by humans, and all students must make physical contact to obtain a one time password or PIN.

5.6 The problem of student identification online and in person

Identification of individuals is achieved by requiring them to produce a secret known only to the individual and the examiner, e.g. student number, password etc. In cases where there is no supervision, students can swap identities and it is impossible to know which student is responsible.

1. For the purpose of examination, one must provide students with a one-off random identity code in each examination. This can be collected immediately before the examination. Each student collects a single key only, by presenting some physical identification and must then not be allowed to pass it to another individual. This makes it unlikely that a surrogate could collect the code for them.
2. After collecting the code, a student could give it to someone else on the way to the examination. Thus the pathway to the examination from the key authorization should be secure.
3. Students must not be able to steal the identity of an earlier student and replay their responses. Secret keys must expire after a short time.

Chapter 6

Contingency plans

What happens in the case of a security breach? This is a matter for College policy, and is yet to be decided.

- For students?
- For staff?

In the case of cheating:

- Expel student?
- Void the grade?

A compromised test can be replaced, within a short time. If too much time passes, fairness is compromised.

Chapter 7

Conclusions and Recommendations

The experiences and analyses lead to a number of straightforward conclusions, concerning grading practice and exam security.

Recommendation 10 (Peer review of courses) *At least two persons should be involved in all stages of a course: all stages need to be peer reviewed by a competent critic. External examiners are not essential, but periodic external review of courses is to be encouraged.*

Recommendation 11 (Exam length) *For reliability, examinations should be as short as possible. The longer an examination lasts, the greater the probability that a student will be able to cheat or abuse the system. Several short tests are thus better than one long test. Students should not feel stress or panic as a result to exam length.*

Recommendation 12 (Checks and controls) *Each examination should contain control questions that exist to test the integrity of the procedure: e.g. questions that cannot be answered correctly (zero grade) or questions that cannot be answered incorrectly. Cheating can be determined by giving each student a unique question.*

Recommendation 13 (Supervised tests) *All courses should include at least one actual examination that has at least partial supervision, as defined above. The identity of students should be monitored physically on arrival at the examination, and students should be issued with one-time passwords or codes (PIN). One should not rely on students' usual usernames and passwords to determine correct identity, since these are regularly swapped.*

Recommendation 14 (Partial supervision) *If complete supervision is not feasible, partial supervision can be used as follows: i) give students the opportunity to work in groups of up to three or four persons, earlier in the course ii) use the knowledge of peer groups to split up and supervise key persons from each group. This strategy will break the most likely route of communication.*

Recommendation 15 (Toilet practice) *Students who have known peer associations should be assigned different colours at the exam entry point. Students of a given colour should be assigned different toilet locations, to prevent message passing. Short exams should not allow toilet breaks, except in emergencies.*

7.1 Expected uncertainties

The following boxes show a somewhat over-simplified summary of the foregoing chapters. They are not a substitute for the detailed discussion in those chapters.

Result 1 (Written examination)
Grade uncertainty: 5%–20% (depending on quality control)
Ease of quality assurance: Medium
Maximum number of students per staff member: 50
Probability of cheating: medium
Probabilty of learning: medium

Result 2 (Electronic examination)
Grade uncertainty: 0–25% (depending on chosen policy)
Ease of quality assurance: Medium to high
Maximum number of students per staff member: limited only by technology
Probability of cheating: medium (requires supervision)
Probabilty of learning: medium

Result 3 (Project work)
Grade uncertainty: 10%–
Ease of quality assurance: Low
Maximum number of students or groups per staff member: 5
Probability of cheating: low-high (difficult to characterize)
Probabilty of learning: medium-high

Result 4 (Grading as incentive to learn)
Grade uncertainty: high
Ease of quality assurance: high
Workload in man hours per student: low – medium (preparation)
Probability of cheating: high
Probabilty of learning: high

Result 5 (Grading as certification of competence)
Grade uncertainty: 5% -
Ease of quality assurance: medium – high
Workload in man hours per student: high (grading)
Probability of cheating: medium
Probabilty of learning: medium

7.2 Grading scale

In humanities subjects, grading involves such a degree of whim and personal opinion

In engineering and scientific disciplines, the uncertainties can be controlled to sufficient degree to speak of an absolute scale of correctness or quality. The Ministry of Educations suggested grade distribution based on student numbers in a given year (see section C.2) is not a fair measurement of student achievement in scientific and engineering disciplines,

since it is based on flawed assumptions: the best and worst students in a class are not guaranteed to be able to pass or fail simply by virtue of their ranking in the class; some answers are accpetable and some are unacceptable. In our Science and Engineering department, a stable scale of quality and correctness can be established. Our chosen interpretation is as follows.

Based on tradition, we draw the pass line at P%.

$$100(1-s) = P \tag{7.1}$$

$$s = 1 - P/100 \tag{7.2}$$

Grade	Width	From	To
A	10s	100	$100 - 10s - \Delta$
B	25s	$100 - 10s - \Delta$	$100 - 35s - \Delta$
C	30s	$100 - 35s - \Delta$	$100 - 65s - \Delta$
D	25s	$100 - 65s - \Delta$	$100 - 90s - \Delta$
E	10s	$100 - 90s - \Delta$	$100(1-s) - \Delta = P - \Delta$
F		$P - \Delta$	0

In order to give students the benefit of the doubt, one can also add the expected uncertainty in grading to their benefit, where appropriate. If we define E as a accepted fail ('vektallsgivende strykkarakter'), the P=35%, or $s = 0.65$, this becomes as in Table 7.1.

Grade	Width	From	To
A	6.5	100	93.5
B	16.25	93.5	77.25
C	19.5	77.25	57.75
D	16.25	57.75	41.5
E	10s	41.5	35
F		35	0

Table 7.1: Table of grade ranges for A,B,C,... given that the grading uncertainty is zero.

Given that the average uncertainty in grading is of the order of 5% for Science and Engineering subjects, we obtain an example grading scheme that can be employed in most instances. See Table 7.2.

Grade	Width	From	To
A	11.5	100	88.5
B	16.25	88.5	72.25
C	19.5	72.25	52.75
D	16.25	52.75	36.5
E	10s	36.5	30
F		30	0

Table 7.2: Table of grade ranges for A,B,C,... given that the grading uncertainty is 5%.

7.3 Further work

The data used in this report are current and accurate, and there is no reason to suppose that any new data will change these dramatically. Nevertheless, it is a part of quality assurance

that, as part of everyday procedure, one verifies the results at every stage.

1. Checklists must be deployed.
2. This document must be maintained.
3. Examination security must be modernized.

(END OF DOCUMENT)

Appendix A

Lectures

This appendix is a personal view on how to conduct lectures.

A.1 Material

Material must be modern and relevant. It should not be 'dumbed down' but rather made 'noble', something to be looked up to. The solution to teaching difficult subjects is not to reduce syllabus or make things easier, but rather to challenge students to take on the challenge by making it appear glamorous.

A.2 Physiological considerations

Time of day is important. The body's production of the hormone cortisol is associated with learning. The rhythm of this hormone is such that levels are high from about 8 am to 14 pm. After this it falls off and learning is much less effective. If levels remain high for a long time, we get chronic stress, can trigger anxiety and depression.

However sleep deprivation can lead to poor attentiveness and is usually the reason why students fall asleep in lectures after lunch. Students should be encouraged to be aware of their sleep requirements (7-9 hours each night). Darkened rooms can also be sleep conducive. Bright light is needed to tell the brain that it should wake up.

A.3 Psychological considerations

Teaching is a fight against *apathy* for many students. Younger students are usually more interested in each other than in dull lectures – so lectures need to entertain, in order to be useful. Older students are more tolerant, but will also appreciate the effort to make lectures entertaining as well as informative.

Moreover, as we mature into adults we introduce more ways to eliminate the thinking process:

- Prejudice - learn the answer to many questions by hearsay and stop thinking about them.
- Rules - we accept and embrace simple rules which we believe give the answer every time, instead of learning to consider every situation anew.

Ideas are intimately connected to language. Many of us never truly believe we understand something until we have expressed it in our own words. Ideas often crystallize only

fully when we have to explain the idea in writing. It helps to encourage students to make written notes.

Most sense input is noise: how do we form structure/language out of such noise? We must construct a relational structure from it. Memory works in this way: by association, not like a hash table. We therefore need to relate important ideas to things which are familiar. It is more important that they be familar than completely relevant, if you just want people to remember stuff.

Examples are no good unless they demonstrate general principles. Presentation of lecturematerial should be structured and logical, but the organization needs to be designed to extract general principles and illustrate them by example.

A good teacher is not one who simply presents the facts, but one who brings them alive and involves the students in the world of those facts.

- By example
- by citing experience and anecdotes,
- By tempting into participation.

The importance of the learning process itself should not be underestimated. It is the experience of learning which burns facts and understanding into our brains both at the conscious and unconscious levels. Creating a satisfactory learning experience involves

- structure and
- patience from the teacher,
- try-and-fail tactics followed by diagnostics,
- participation,
- regular success and praise for the students.

A treasure trail of learning can stimulate students to learn. Make learning a puzzle for them to solve.

A.4 Problems and exercises

The construction of problems is central to the learning process. Each key principle or point from lectures should be accompanied by problems which demonstrate or explore the idea. Problems should be a progression, always motivated clearly.

- The first of a sequence of problems should be trivial and confidence building.
- Then more difficult examples of the *same* problem should be given, in order show a simple thing in a difficult context.
- It should be clear to students that the problems they are solving are useful and they should know what they have achieved when they have done the problems.
- If possible, problems should result in a feeling that they have advanced their knowledge.

A.5 For the teacher

1. Plan the start of your lecture to capture attention.
2. Explain what the students should expect to learn from this lecture.
3. The main purpose of a lecture is to inspire students to study on their own. Students will only learn a handful of things in each lecture, so pick the points to be emphasised carefully.
4. Be active on your feet, animated, project energy. Avoid habits like pacing back and forth, saying ummm and aaahh. These are distracting and send people to sleep. Try to vary the tone and level and style of speech for dramatic content. Show passion!
5. Don't be afraid to criticize something in a lecture, but always give a reason. e.g. "The C++ stream library is just awful – a step backwards from C, because it takes 3 times the coding to say what you want and the result is completely unreadable". As long as you are authoritative in your criticism, this will be humorous. If you show "uncertainty" or "indecisiveness" students perceive this as weakness, and you will lose their respect.
6. Ask rhetorical questions without picking on people: what if, imagine if... Let the students answer if they want, but don't waste time waiting for an answer that usually doesn't come.
7. Explain often "This is very interesting because...." and give a reason Don't assume that they understand why something is interesting or relevant. It is very important to explain why things are interesting. Most people do not have the experience or perhaps insight to see it for themselves. Explain the implications!
8. Keep them busy in the lecture, writing and thinking or grab their attention with something unexpected but relevant.
9. Be sure to provide sufficient breaks:
    - Pauses to think (not too long) during a lecture.
    - Breaks to stand up and move around after each period (about 45-50 minutes).
    - Break at an appropriate time – not just at an appointed time.
    - Break if students appear to be losing concentration.
10. Don't let students close their eyes, talk, or read the newspaper in class. They need to demonstrate a respect for the classroom process (other students and the teacher).
11. Students should leave every lecture feeling that they have learned something and with something to think about.

Appendix B

Writing multiple choice questions

Students should have to think carefully to select the right choice. They should have to engage not just recognition faculties, but analytical ones.

The most obvious method for automated evaluation is the multiple choice (MC) question. Multiple choice questions must be answered by an individual, but can be used to stimulate discussion within a group (see fig. B.1). MC evaluation can be done independently of humans, and it is therefore not subject to simple grading errors. It scales easily to large numbers of students and it thus a very 'cheap' solution. The difficulty here lies in crafting the questions.

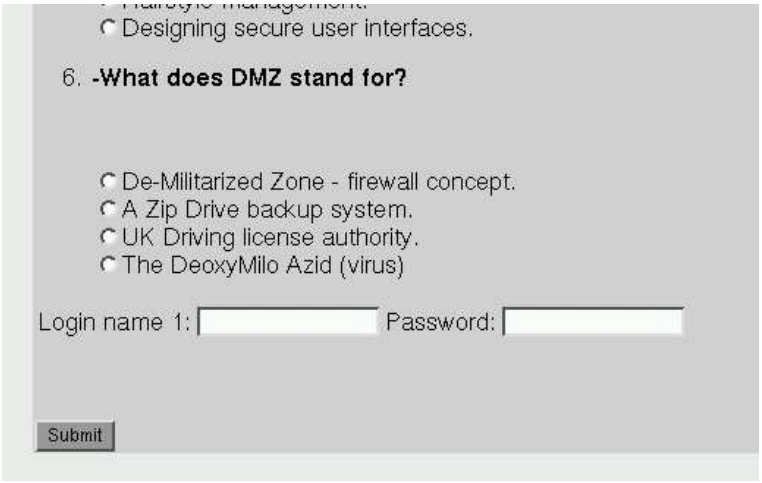

Figure B.1: A multiple choice test. The user submits his or her choices by 'signing' with a login name and password.

If one thinks in traditional terms, i.e. that the purpose of an MC test is to find out what a student knows, then MC questions which test knowledge of facts are easily criticized: facts can be looked up in books, and whether or not a student remembers them is more a function of memory than competence. However, MC questions can be used to force students to read reference material or review work which they are already supposed to know. They can also force students to learn from past errors, through repetition.

MC questions which test students ability to carry out a task, or their understanding of a topic can be very useful. This is often called a psychometric evaluation, or aptitude test. The student is presented with a number of questions, with a number of possible answers and is required to select those answers which are correct, or incorrect. There are several MC styles: i) Only one correct answer; ii) Several correct answers (not recommended);

iii) Only one wrong answer (not recommended), and so on. For instance, questions which begin 'Which of the following is true...' are less confusing than questions which begin 'Which of the following is not true...'; the latter are often ill-conceived and students find them confusing.

In general, MC questions are very fragile with regard to phrasing. A tiny ambiguity can change the perceived meaning of the question; unfortunate choice of wording can lead to misunderstanding, or even the rejection of the question by the students as 'silly' or invalid. This places a burden on the author of the questions, and tests on willing subjects are usually required to iron out problems.

Examples of a question in which students have to think in order to select the correct answer. The answers are sufficienly similar that they require careful reading, and effort to distinguish them:

1. -Would you recommend buying a switch to improve password security?
   (a) No because it is an expensive way to only slightly improve security.
   (b) No because it slows down traffic in busy organizations.
   (c) No because it makes it easier to collect passwords centrally.
   (d) Yes because it provides complete privacy of traffic between individuals.
   (e) Yes because switched traffic is encrypted.
   (f) Yes because a router is the correct solution to this problem.
2. -A digital signature
   (a) Can only be added to a message by the owner of a private key
   (b) Can only be verified by the owner of a private key
   (c) Can only be removed and decrypted by the owner of a private key
   (d) Can only be accomplished with symmetrical encryption

From a security viewpoint, one is interested in preventing students from simply copying answers from one another. There are two possibilities here. Questions can be picked at random from a pool, so that each student receives a difference set of questions. This might not be desirable however, or even possible without a sufficient pool. The order of answers, within a question can also be randomized, so that students cannot copy directly from one another, at least without the responses passing through their brains. We have seen evidence that students have tried to blindly copy answers and were foiled even by this simple tactic, however, errors in entering the correct answer were also caused by a general lack of respect for the procedure.

If students do their research properly, there is no reason why they should not always get one hundred percent on each MC test, but in the philosophy of grading as an incentive, this is not a problem. The point is that actual learning was achieved, and the grade is a reward for doing this work, rather than being an ephemeral attestation of competence.

Experience shows, however, that students have two problems, both of which might be computer related:

- Carelessness - they do not respect the quality of the input they enter into a computer program. They are used to clicking and 'undoing' poor decisions.
- Guesswork - they tend to guess rather than user their analytical ability to solve a problem – they are used to being able to solve most problems on a computer by trial and error.

Appendix C

Summary of data from various types of examination

C.1 Mathmatics exam with external check

The following data are taken from previous Norwegian examinations in statistics. This is an right/wrong answer examination. The data indicate how much deviation of opinion there was between the course lecturer and the external (control) examiner.

Kurs:	No. students	No. $\Delta > 2.5\%$	No. $\Delta \sim 5\%$
F202 v 2002	17	6 (p=0.35)	1 (p=0.06)
F202 v 2001	19	3 (p=0.18)	0 (p=0)
Fysikk Haugesund	124	17 (p=0.14)	0 (p=0)
Sommereks ma1	91	30 (p=0.33)	8 (p=0.9)
Statistikk Oslo 2000	276	60 (p=0.22)	9 (p=0.13)
Konte 2002	34	11 (p=0.32)	5 (p=0.15)
Statistikk Oslo V2001	203	55 (p=0.25)	16 (p=0.08)
Average		$\langle p \rangle = 0.26$	$\langle p = 0.19 \rangle$

We take these numbers as typical of what is achievable in human grading of definite answer questions. The expected error is thus of the order

$$\langle \Delta \rangle \sim \frac{\langle p(2.5) \rangle 2.5 + \langle p(5) \rangle 5}{\langle p(2.5) \rangle + \langle p(5) \rangle} \sim 3.6\% \tag{C.1}$$

C.2 Data on grade uncertainty due to MoE's suggested grading scheme

Consider the data in fig C.1. The Ministry of Education has suggested that grades should be partitioned according to percentiles of the number of students:

$$(A, B, C, D, E, F) = (10, 25, 30, 25, 10). \tag{C.2}$$

The idea is that, by partitioning the yearly grades of students actually participating, rather than comparing to a static achievement scale, one takes into account the variable nature of external factors from year to year. However, this does does not lead to any reduced uncertainty, on the contrary. An attempt to fit this distribution to previous grade distributions over the course of a number of years, shows a wide discrepancy. Measured according to the old Norwegian grading scheme 1,0-6,0, the uncertainty is of the order of 0,5 as compared to 0,1 for the best-case static scale for a strictly graded test.

Hvor store utslag gjør selve oppgavene på karakterfordelingen.

En studie av fagene Matematiske metoder 1, DMLA, Statistikk og Fysikk viser følgende overgang fra gammel til ny skala dersom vi følger fordelingen 10, 25, 30, 25, 10 % av studenten som har bestått skal ha karakterene A, B, C, D, E

	Grenseverdier for karakterene ved en overgang fra gammel til ny karakterskala					
Kurs\år	A	B	C	D	E	# studenter
Matematiske metoder 1 1999	1.0 - 2.0	2.0 - 2.9	2.9 - 3.2	3.2 - 4	- 4.25	
2000	1.0 - 1.8	1.8 - 2.8	2.8 - 3.3	3.3 - 4	- 4.25	
2001	1.0 - 1.4	1.4 - 2.4	2.4 - 3.4	3.4 - 4	- 4.25	
DMLA 1999	1.0 - 1.2	1.2 - 2.0	2.0 - 2.6	2.6 - 3.9	- 4.25	
2000	1.0 - 1.4	1.4 - 2.4	2.4 - 3.0	3.0 - 4	- 4.25	
2001	1.0 - 1.7	1.7 - 2.9	2.9 - 3.4	3.4 - 4	- 4.25	
Statistikk 1999	1.0 - 1.2	1.2 - 1.9	1.9 -2.5	2.5 - 3.7	- 4.25	
2000	1.0 - 1.7	1.7 - 2.3	2.3 - 2.8	2.8 - 3.7	- 4.25	
2001	1.0 - 1.7	1.7 - 2.3	2.3 -2.7	2.7 - 3.6	- 4.25	
Fysikk 1999	1.0 - 2.0	2.0 - 2.7	2.7 -3.2	3.2 - 3.9	- 4.25	
2000	1.0 - 1.5	1.5 - 2.4	2.4 - 2.8	2.8 - 4	- 4.25	
2001	1.0 - 1.8	1.8 - 2.7	2.7 - 3.2	3.2 - 3.9	- 4.25	

Konklusjon: Karakterene A og B er relativt stabile mens B-C, C D og D-E har stor spredning. Denne usikkerheten er mye større enn usikkerheten knyttet til selve sensureringen. Hvis man skal jobbe med kvalitetssikring av karakterer er dette det punktet man vil ha mest igjenn for å forbedre.

Figure C.1: Statistics showing the uncertainty in grade, on fitting several years of data to the grade distribution function proposed by the Ministry of Education. The data indicate that grades cannot be meaningfully fixed to a better accuracy than ± 0.5 in the old grade scale.

The sources of uncertainty are the variability of the exercises, in addition to the student performance and other factors. Since this uncertainty is greater than the pure grading uncertainty noted above, this tends to suggest that a large part of the grade variation is due to quality assurance factors in teaching and examining.

The Ministry of Education's grade partitioning is based on the idea that student ability is normally distributed from year to year. This is clearly not true, and thus there is a large discrepancy. Moreover, the suggestion that only 10% of students can succeed maximally in any class implies that any attempt at quality control is in vain: no amount of procedural quality assurance will allow us to change this distribution to make anything other than a normal distribution.

C.3 Peer review

From courses on computer security. 120 students in 2001 and 150 students in 2002, in group work, groups of three, with three peer review reports. The deviation in grading of any student's work was:

$$\langle \Delta \rangle \sim 4.7\% \tag{C.3}$$

using guidelines for marking. The maximum deviation was

$$\Delta_{\max} = 19\% \tag{C.4}$$

C.4 Questionnaire to students and staff to gauge uncertainties

To staff:

1. For each of your examinations, please estimate the percentage error in grading the exam, by working out the *average difference* between your grade and the external examiner's grade:

$$\Delta = \frac{\text{Average difference in grade}}{\text{Total grade}} \tag{C.5}$$

2. Does the majority of the exam questions have right or wrong answers, or are there value judgements involved that must be graded by 'feeling'?

3. If the exam is anonymous, how often can you guess the identity of the student? Has this ever affected the grade you give?

Appendix D

Electronic Test Types

- Multiple choice: one correct answer amongst many.
- Active area: click on the correct part of a picture.
- Fill-in questions: student has to fill in blanks in a form.
- Drag’n’drop: like round peg in round hole tests.
- Order/sort question: student must rank or sort a list.

Appendix E

Prevention and forensic detection of cheating in E-exams

The test:

1. Give students PIN codes as they enter the exam, after verifying their identity.
2. Give a short test that is timed.
3. Give them a receipt code on submitting the test.
4. Make students leave the controlled examination area and register their PIN code immediately at a special location, within a few minutes of finishing. (This prevents students from communicating with others.)

After the test:

1. Determine social networks/groups. Which students work together and have other social ties? Classify these groups into overlapping sets.
2. Obtain a time series list of submissions with start and finish times.
3. Look for correlated scores in these groups, e.g. students with 100% starting from the first student who is sacrified.
4. Cheating is indicated by an attempted covert communication between users in the group, e.g. from printer log records, or E-mail records.