TEACHING AND LEARNING SERVICES

UTQ Assess Reader

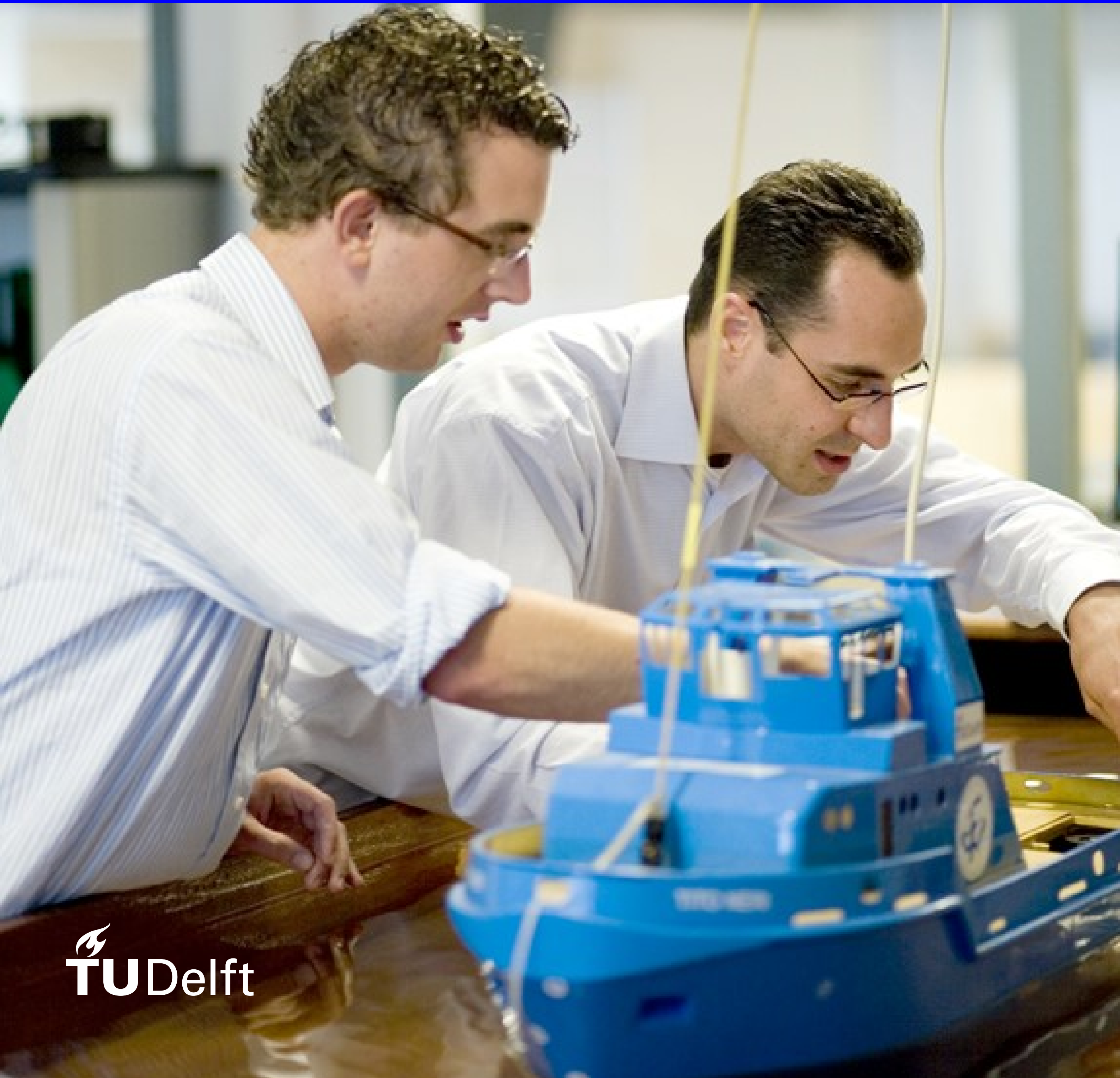


TABLE OF CONTENTS

NOTE: This document is for internal use only and may not be distributed without permission of the owner

Compiled and edited by Lisette Harting, OC Focus, TU Delft, October 2017
Last update: Erna Engelbrecht & Lisette Harting (Teaching & Learning Services, TU Delft), August 2018, November 2018, February 2019, April 2019, March 2020.

INTRODUCTION

In this reader, you will find guidelines on how to develop good quality assessments. We will use this reader during the UTQ module ASSESS, and you are encouraged to use it as a reference for your own course.

The purpose of assessment is not only to attach a grade to the students' level of knowledge and performance. It should also enable and increase learning. By combining formative assessments, feedback and summative assessments, you can steer your students' learning behaviour in the most optimal way. Note that the word '**assessment'** can refer to the collection of exam/assignments/projects within a course (as in 'the assessments of a course'), as well as an individual exam/assignment/project.

During the course, we will guide you through the most important steps of the assessment cycle (see

Figure 1). The following steps will be covered:

Development of an assessment plan (1) in which you draw up a plan of how you will combine formative and summative assessments in your course, and how all this will lead to a grade.

The quality requirements for assessment and to the rules, regulations and assessment policies that apply to your course. We will also discuss how you can use the test results to measure your students' mastering of the learning objectives (2), estimate test quality (3), and to (re)calculate the test grade (3).

How to design assessments, or improve existing ones, using three steps: 4) Develop or improve the blueprint of the assessment; 5) Develop or improve the assessment instructions; and 6) Develop or improve the assessment criteria.

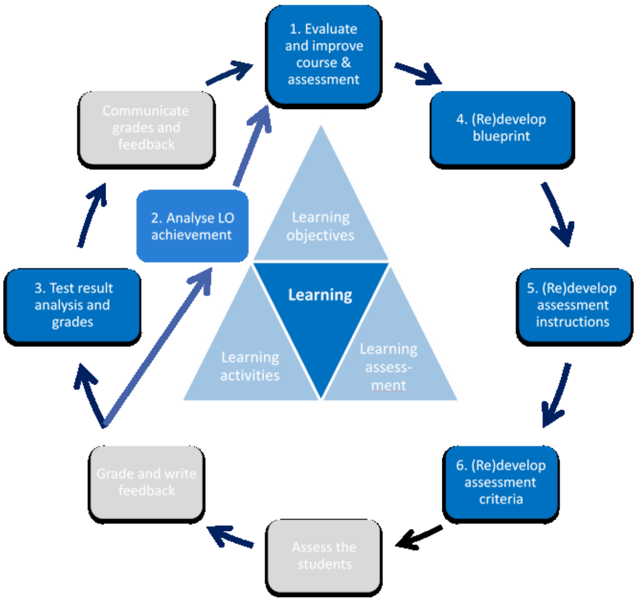

1. Evaluate and improve course & assessment
4. (Re)develop blueprint
5. (Re)develop assessment instructions
6. (Re)develop assessment criteria
Assess the students
Grade and write feedback
3. Test result analysis and grades
2. Analyse LO achievement
Communicate grades and feedback
Learning objectives
Learning
Learning activities
Learning assess-ment

Figure 1. Assessment cycle

Compiled and edited by Lisette Harting, OC Focus, TU Delft, October 2017
Last update: Erna Engelbrecht & Lisette Harting, August 2018, November 2018, February 2019, April 2019, March 2020.

CHAPTER 1: ASSESSMENT PLAN

An assessment plan contains detailed information about the constructive alignment of your course assessments, and how the assessments contribute to the course's final grade. In this chapter, it is explained how to construct, analyse and improve such an assessment plan for constructing or improving the assessment of your course. In this chapter, we will look at the assessment plan in detail, and give an example of what it could look like.

1.1. Assessment plan, analyses and characteristics

To give a good overview of the constructive alignment of your course assessment, include an *assessment overview* (a tabulated summary of the assessment plan). See, for example, Table 2. This assessment overview can be included in your Brightspace course for your students to see what assessments they can expect. Because it is a summary of the entire assessment plan, we use the assessment overview to get insight on the following at a glance:

- Constructive alignment of assessment methods with learning objectives;
- Alignment of formative and summative assessments with feedback;
- Grading methods;
- Timing of assessments and feedback.

It is recommended to include the elements listed in Table 1. These elements will give insight on the level of validity, reliability, transparency and feasibility, in your course assessment. Refer to Table 2 for an example.

Depending on your course, you could leave out certain elements that are less relevant or that would be difficult to summarise in an assessment overview table, and describe them only in the running text of the assessment plan.

In the table, you summarise **all formative and summative assessment** in your course. Summative assessments test how well students master the learning objectives. Summative assessments may be classic written exams, digital exams, assignments that students perform at home or during a computer lab, performance, presence or attitude during for example a project, lab, excursion or class. Summative assessments usually lead to a grade (1-10), or a pass or fail decision.

Table 1.
Assessment plan characteristics, divided into three assessment plan analyses.

General	
Assessment name	Descriptive name of all assessments (formative and summative)
1. Assessment method alignment	
Assessment method	Examples: midterm exam, homework assignment(s), project, presentation. It is important for the method to be aligned with the learning objective.
Individual / group	In case of a group: group size
LOs	List of the assessed learning outcomes
2. Alignment of assessment types	
% of final grade	Percentage of the final grade that each assessment determines (formative assessment is 0%)
Grade type	How the assessment is evaluated (grade (1-10), points, pass/fail, feedback only, etc.)
Feedback on assessment outcome	Type, focus and communication medium of the feedback. Examples: rubric (or 'grade only', or group feedback form), focussed on the final paper, communicated via Brightspace.
3. Regulation compliance	
Minimum grade	What minimal grade the student needs to achieve in order for the grade to count for the final grade (see TER)
Deadline or date of assessment	Completion or scheduled dates
Grade and feedback due date	Timing/dates of release of grades and feedback. In the case of formative assessments, there should be enough time available for the students to improve their work/knowledge before the summative assessment.

Formative assessments are assessments that usually do not contribute to the grade of the course. Here, students should receive feedback on how well they master the learning objectives. This can be done by giving teacher/TA feedback, automated computer feedback or peer feedback. The resulting feedback is focussed on criteria that cover the tested learning objectives and the assignment is at the same level as the summative assessment. This way, the formative assessment prepares students for the summative assessment.

Let us look at the assessment overview and plan in more detail.

Table 2.
Example assessment plan

Assessment name (assessment type)	Assessment method	Individual or group	LOs	% of final grade	Grade type	Minimum grade	Deadline/ date of assessment	Grading method	Date of announcement of grade/ feedback	Feedback on assessment outcome
ECG analysis (assignment)	Report, code, presentation	Group	3,4,5,6	20%	Grade	5 for the weighted average of two assignments	End of week 4	Rubric	End of week 5	Rubric with a tip and a top, feedback is focused on EEG analysis assignment and on the exam.
EEG analysis (assignment)	Report, code, presentation	Group	5,6,7,8	30%	Grade		End of week 9	Rubric	End of week 10	Rubric with a tip and top, focused on the exam.
Excursion Medical Company	Attending the excursion	Group	3-8	0	Pass-fail	pass	Week 5	None	Immediately after the excursion	NA
Practice exam	2 open questions with 4 sub-questions, 40 multiple choice questions with 3 options each	Individual	1-4, 7	0%	NA	NA	Start of week 10, in class	Answer model	Immediately after the practice exam	Exam and model answers are on Brightspace, including references per sub-question to page numbers and exercises in the book. Students can ask questions in class after the exam.
Exam	2 open questions with 4 sub-questions, 40 multiple choice questions with 3 options each	Individual	1-4, 7	50%	Grade	5	End of week 11	Answer model	Week 13	Debriefing after exam. Exam and model answers published after exam on Brightspace, see practice exam.

1.2. Describing the assessment plan

Here is an example of a summary of the main characteristics and considerations of an example assessment plan, based on the assessment overview on the previous page:

1.2.A. Minimum grade

The reason that there is a minimum grade for the assignment, is that this is the only place where LO5 and LO6 are summatively assessed. However, the grades of the two assignments can compensate each other. Students have the biggest problems with mastering LO5 and LO6. Since both the assignments contain these LOs, and because the second assignment has a higher weight, students can use the feedback on the first assignment to improve on LO5 and LO6 in the second assignment. Therefore, it is fair that they can compensate the assignment grades, since they partially measure the same LOs and that redoing assignment 1 would be partially redundant and unnecessarily increase the workload for students (doing an extra assignment) and lecturers (grading these assignments).

1.2.B. Retake for the excursion

Since the excursion is mandatory for finishing the course, but there may be cases where students are not able to visit the company, students who have a valid reason not to attend the excursion (to be determined by the study advisor) are allowed an alternative, for example, writing an essay on the company visited by the rest of the class.

1.2.C. Grade valid after the end of the course

Since the assignments change every year and reflect the state of the art developments in the field, students cannot keep the grade the next year. Furthermore, the assignments are group work, and if students passed the group assignments, but failed the course because they had a low grade on the exam, they may not have contributed enough to the project after all, since they apparently lack some knowledge and skills. Finally, speaking from my own experience: In previous years, students did not have to retake the assignments. However, students who did not get a pass for the assignments before taking the exam, almost never passed the exam. To be able to pass the course, students would need a second chance to complete the assignment successfully.

Since the excursion is the same every year, students are not required to go on the excursion a second time.

1.2.D. Feedback

For the **assignments**, students get their grade, rubric and *tip & top* one week after the deadline of the assignment. We have enough TAs and lecturers to do this. Furthermore, students will get this feedback before starting the next assignment and one week before the exam, and are encouraged to use this feedback to work on assignment 2, and to study for the exam. The *tips & tops* are only focussed on the next assessment so that the students can actually apply the feedback.

Students will be advised to take the **practice exam** at home, once they think they are well prepared. If students get stuck on a question, they can use the hints on a hint-form, which will refer them to a page or formula in the book (the exam is an open-book exam), or to related exercises, which they can use to get to the answer or practice more. After finishing the practice exam, the students can compare their answers to the model answers. To make sure that students realise that there are more correct ways to get to the correct answer, multiple answer routes will be included in the model answer.[1]

The model answers will be published on Brightspace. In these model answers, each model answer of a subquestion will have a reference to a page or formula in the book and to related exercises, so that students can study and practice that part, in case they will take the resit.

Directly after the **exam**, students are invited to a neighbouring lecturing hall, in which the lecturer will discuss how the questions could have been answered. The lecturer will emphasise that the goal of this meeting is to enable learning after the exam, not to discuss the quality of the questions, since students will be able to inspect their work and file complaints in another meeting, after the grades have been announced.

Just like for the practice exam, the model answers will be published on Brightspace with references to the book and exercises, so that students can study and practice, in case they will take the resit.

Of course, circumstances in your course are different.

[1] Some lecturers choose to publish the real answers from several students who used different approaches that led to a correct solution. This will stimulate students to find their own creative solutions.

1.2.E. Minimum grade implies retake

Whenever a minimum grade is present, it is recommended to grant students a retake, or enable them to deliver a new version of project reports. The reason for this is to diminish the number of assessment hurdles for students, since grades are not perfectly reliable and erroneous grading may keep students from progressing with their studies. That is why there is a retake for the assignments as a whole, and a retake for the exam. Another reason to average the grade of the assignments, is to lessen the workload for the teaching staff. If a crucial learning objective is only assessed in a single assignment, it would be good reason not to average the grade, and instead to require students to earn a minimum grade for a single assignment.

1.3. Assessment methods and Constructive Alignment

This UTQ course is centred around constructive alignment of assessments. For your students to complete your course, they should demonstrate their knowledge and skills in some way or another. They demonstrate this by completing the summative assessments that you set for them. Once they have completed an assessment, you then evaluate/grade them based on certain predefined criteria. These criteria should be based on the learning objectives of the course.

Now, to enable your students to complete these summative assessments, you will have provided them with various learning activities to enable them to prepare. This might include course content, excursions, lectures, workshops, formative assessments etc. Lastly, you close the loop by checking that absolutely everything in your course (whether it is content or assessments) will enable your students to reach the learning objectives for the course. If so, your course is constructively aligned.

All assessments should cover at least one learning objective, otherwise there is no point in including it in your course. If assessments do not aim towards students meeting the learning objectives for the course can be considered redundant.

In the following two sections, we will discuss how the choice of assessment method as well as the balance between formative and summative assessments will influence the constructive alignment of your course. This text is adapted to the TU Delft situation from (Dunn, Selecting methods of assessment, 2018).

1.3.A. Choosing the right assessment methods

Assessment methods are, for example, written tests, presentations, and projects. It is important to select the right type of assessments for students to show whether or not they have reached the learning objectives. For example, if you want to assess students' communication skills, you would rather have them do presentations than a multiple-choice test.

The main reason to choose one assessment method over the other is that it enables you to get a valid measure of how well a student masters a learning objective. The assessment should be authentic for you to be able to assess what you should be assessing.

During an assessment, students should be able to demonstrate their capabilities, unhindered by the lack of experience with an assessment method. If you use an assessment method that students are not trained in (for example oral exams, group assignments), the assessment method should not prevent students from maximum performance. For example:

When the learning objective is to (orally) explain and defend design choices for a given case, it is okay to use oral exams, if and only if students can practice orally with this during the course, and receive good quality feedback on the criteria that they will be assessed on, while practicing (formative assessment). And if all measures have been taken to ensure validity, reliability (assessor objectivity, as well as creating a safe atmosphere to enable maximum student performance), and transparency, since these quality requirements for assessment are more easily violated than using other assessment methods.

Keep in mind that the learning objectives contribute to the overall aims of the programme, and may include the development of (inter-)disciplinary skills (such as critical evaluation or problem solving) and support the development of vocational competencies. Ideally, this should be planned together with the relevant colleagues so there is an purposeful assessment strategy across a degree program.

To motivate students to do the assessments and to do them well, it is important to **validate** why any particular assessment type was chosen. This works best if the assessment is ***authentic***, i.e. if they will perform the activity during their working life, or otherwise during a follow-up course. This will make the assessment much more **relevant** for your students, and will also help them decide if they want to pursue a career where that type of activity is common.

Nightingale *et al.* (1996) provide eight broad categories of learning outcomes which are listed here. Within each category some suitable methods are suggested.

Note that oral exams are not included, since they are only advised when the learning objective requires it, for example ‘being able to defend one’s ideas within a research team’.

Table 3: Categories of learning outcomes (Nightingale et al, 1996)

Thinking critically and making judgements	
Developing arguments, reflecting, evaluating, assessing, judging	- Essay - Report - Journal - Letter of advice - Case presentation for an interest group - Committee briefing paper for a specific meeting - Book review (or article) for a particular journal - Newspaper article for a foreign newspaper - Comment on an article's theoretical perspective
Solving problems and developing plans	
Identifying problems, posing problems, defining problems, analysing data, reviewing, designing experiments, planning, applying information	- Problem scenario - Group Work - Work-based problem - Draft a research bid to a realistic brief - Analysis of a case - Conference paper (or its structure plus annotated bibliography)
Performing procedures and demonstrating techniques	
Computation, taking readings, using equipment, following laboratory procedures, following protocols, carrying out instructions	- Demonstration - Video (write script and produce/make a video) - Poster - Lab report - Illustrated manual on using the equipment, for a particular audience - Observation of real or simulated professional practice - Role play
Demonstrating knowledge and understanding	
Recalling, describing, reporting, recounting, recognising, identifying, relating and interrelating	- Written examination: - Open questions - Essay questions - Short answer questions - Closed-ended questions: o True/false o Multiple choice - Paper-based or computer-aided - Essay - Report - Comment on the accuracy of a set of records - Devise an encyclopaedia entry - Write an answer to a client's question
Designing, creating, performing	
Imagining, visualising, designing, producing, creating, innovating, performing	- Portfolio - Presentation - Projects - Performance
Accessing and managing information	
Researching, investigating, interpret-	- Annotated bibliography

ing, organising information, reviewing and paraphrasing information, collecting data, searching and managing information sources, observing and interpreting	- Project Dissertation - Applied task - Applied problem
Communicating	
One and two-way communication; communication within a group, verbal, written and non-verbal communication. Arguing, describing, advocating, interviewing, negotiating, presenting; using specific written	- Written presentation (essay, report, reflective paper etc.) - Oral presentation - Group work - Discussion/debate/role play - Participate in a 'Court of Enquiry' - Presentation to camera - Observation of real or simulated professional practice
Managing and developing oneself	
Working co-operatively, working independently, learning independently, being self-directed, managing time, managing tasks, organising	- Journal - Portfolio - Learning contract - Group work

Please note that these suggestions are not focussed on engineering education, and you as a lecturer and as an expert in your own field will probably have other ideas for assessment methods that are more authentic in your situation. It will hopefully expand your view on the possibilities of assessment methods beyond the classical closed-book exams

1.4. Choosing between open and closed-ended questions

In general MCQs in which students have to demonstrate understanding, are very useful in a classroom setting were students can discuss their answers. This can deepen their understanding and analytical skills. However, you might consider the above for your decision on using a summative MCQ for assessing the learning objectives of your course.

If you choose to use closed-ended questions, such as multiple-choice questions (MCQs) in an exam, keep the following advantages and disadvantages in mind:

1.4.A. Advantages

MC questions that test lower levels of Bloom, can be answered quickly. Therefore, you can include many questions, which can increase validity and reliability.

The grading can be very fast, and will automatically provide you with data for doing item analyses.

It is possible to test higher cognitive levels of Bloom, but more time need to be spent on creating these questions. A good idea is to use case studies which the students have to analyse, and then base your questions on the cases.

1.4.B. Disadvantages

Generating MCQs takes a lot of time and should not be seen as an easy way out. A lot of care need to go into developing really good questions, and building a large enough library of questions can take a while. Keep in mind, for example, that all distractors must be equally probable.

If you want your students to recall facts ('remember' level of Bloom), do multiple choice questions measure whether the students can recall the facts, or do MCQs merely measure whether your students can recognise the correct answer between false answers? Do you measure whether your students will be able to produce the answers by themselves?

The same holds for higher levels of Bloom, which has as an extra problem that students will most likely need more time to answer each question. Since you will need quite some multiple choice questions in order to develop a reliable test, this might be problematic.

For MCQs that need a lot of thinking steps, like ones with calculation or difficult case studies, generally no partial credits are given to partially correct answers, whereas for equivalent open questions partial credit would be given. Please note that it is possible to give partial credits to partially correct answers in Contest (paper-based MCQs), and probably also in other software. However, this will influence the guessing score.

On the other hand, the student might have guessed the correct answer, without having studied the subject. In open questions, the student would probably have gotten 0 or very few points.

The latter two points will create noise in the grade, which will make the grade less reliable. That is why you will need more questions for multiple choice questions than for open questions in order to construct a reliable exam (see 1.28.D, 'Number of exam questions').

1.5. Digital assessment tools

There are tools that help you to **grade paper exams online**. Some of these tools allow you to divide the grading work amongst graders, grade anonymously, grade per question, and grade simultaneously with your fellow graders.

- Zesje (open source, Latex based, no possibility yet to distribute questions amongst graders)
- ANS Delft (availability depends on the licences of the faculty)
- Work2grade (TBM, Pieter Bots)
- Grasple (Annoesjka Cabo; currently only available to the Maths Department, but could also be used for testing statistics).

There are also **peer evaluation/feedback** systems:

- BuddyCheck (to improve behaviour and group dynamics, follow-up off Scorion)
- FeedBackFruits
- Turnitin (login directly to the website for peer review, instead of Brightspace)

Reminder: the examiner is responsible for giving the grades!

You can contact Brightspace Support (Brightspace@tudelft.nl) if you need more information, or if you want to use other tools for assessment.

For **summative** and **formative digital exams**, Maple TA is the recommended solution at the moment. This allows you, for example, to use adaptive testing, where a student is presented with a follow-up question based on their performance on a previous. For more information and support, contact digitalexams@tudelft.nl.

1.6. Balance of formative and summative assessment, and feedback

It is important to include a balanced combination of formative and summative assessments in your course. While summative assessment is used to collect evidence on the extent to which students master the learning objectives, formative assessment is meant to steer learning. Let us look at this in more detail.

The main difference between formative and summative assessment is that formative assessment does not contribute (significantly) to the final grade of the course. For formative assessments, the students should focus on their own learning (Garfield & Franklin, 2011), make mistakes and experiment with new ideas without any significant consequences for their final grade. This is **assessment for learning**. Furthermore, the lecturer can use the information on student performance to adjust the course to the need of this particular group of students.

Formative assessment has been shown to have the following positive effects (Cauley & McMillan, 2010) (Shute, 2008) (Wiliam, 2011), for example:

- Pointing out misconceptions and allowing them to be corrected;
- Providing valuable information for the adjustment, or improvement of instruction;
- Allowing students to be more actively engaged in their own learning and increasing commitment.

Formative assessments have to meet certain conditions to enable successful completion, for example:

- The lecturer needs to believe in the value of each formative assessment, set high expectations from the start, and follow a consistent approach throughout the course;
- The purpose and reason for each formative assessment have to be explained to students, as well as the goals and the evaluation criteria;
- Students have to want to be actively involved in their own learning;
- Feedback must be timely (as soon after completing the assessment as possible) and contain information about how the student is doing, where the student is going and what (s)he still needs to do to get there.

In short, formative assessments are all types of planned assessments during a course that are non-binding (no grade attached) and in which students participate voluntarily in order to receive feedback on their learning process.

Watch this video that explains four characteristics of effective feedback:
https://www.youtube.com/watch?v=Huju0xwNFKU

So, if the students are not graded for these assessments, how do you get the students to complete them?

- Manage the students' expectations at the start of the course (let them know what they can expect and what will be expected of them);
- Make the formative assessments their gateway to performing well on the summative assessment (it has to be worth their time coming to class);
- Clarify what kind of feedback students can expect and how this will help them;
- You could make the formative assessment fully optional, so that students who prefer another way of studying can choose to choose to do so;
- Coordinate the assessment methods, deadline, and bonus point arrangement with other courses in the programme that are running that period and year, so that the assessment activities do not clash;
- Adjust the type of feedback to the year the students are in.

The most important thing is that you offer students the opportunity to get feedback on their performance, per learning objective, at the level of the summative assessment, before grading them. You can do this, for example, either by writing general feedback, personalised feedback, using rubrics, or a combination of these.

One purpose of **giving feedback** to students is always to steer their progress. This means that feedback should at least answer the following questions for the student:

- Feed up
  - Where should I go to?
  - What is the required level?
- Feedback
  - Where am I right now?
  - What is my current level?
- Feed forward
  - What is the first step I need to take in order to get closer to my goal?
  - What can I do now to improve your level?

The student should know what the goal is, and why it is important to reach this goal. This should also be made clear before they start with the assessment.

Another purpose of giving feedback is to help the **lecturer** understand how students are doing in the course and what they will still need (from the lecturer) to reach the learning objectives. He or she can then use this information to make adjustments to the course while it is still running, allowing for better learning results.

Having **feedback mechanisms in place during group work** assignments is very important. If your students first complete the summative assignment and only then receive feedback, it is too late to improve their learning objective achievement. Instead, have your students for example give feedback on each other's' work half-way through, or at certain milestones in the project. If there are problems with any of the performance areas, they will still have time to correct these, instead of reaching the end when it is too late to address any issues.

Furthermore, the specificity, practicability and **respectfulness** of the feedback can be ensured by using the **'Observation, Result, Advice'** structure in the formulation of your feedback, no matter whether the feedback is positive or focused on improvements:

STRUCTURING FEEDBACK
Observation
What did you observe? Start with 'I noticed.../I observed that.../In question 2 I see that...' and describe your observation. Your observation should be based on evidence.
Effect
What was the result? Describe the effect it had on you, or the effect it might have on other readers/listeners/professionals.
Advice
Give a concrete hint on how to improve or do things differently, or (if correct) encourage the student to maintain this behaviour.

By following these steps, you will both indicate **why** (i.e. **validations**) and **how** the improvement could be made, in an objective way that is specific, respectful, and actionable.

Here are three examples of how to apply the 'Observation, Result, Advice' structure:

Feedback example presentation:

1) I noticed that during the presentation, you talked quite fast.
2) For me, this made it hard to follow your talk.
3) Maybe you could practice on speaking slower. If you are talking fast because you are nervous, you could try doing some breathing exercises before the presentation. There are plenty of examples on the Internet.

Feedback example code:

1) I noticed that you did not use section headers or comments.

2) This made it very difficult for me to understand what part of the code is doing what, and it took me a lot of time to understand it. As a result, your grade for 'code readability' is low.
3) You can improve your code's readability by using logical section headers and adding comments. You can find some examples on page 13 of the book.

Feedback example report:

1) I could not find a critical discussion of your research method in your research paper.
2) Therefore, I could not check how you have taken into account the limitations of your method in your conclusions. As a result, you have a low grade for 'reflection on methodology'.
3) Please add a critical reflection on your methodology in your discussion. You can have a look at the example research paper on Brightspace, which has a good example of what is expected.

As you can imagine, giving such comprehensive feedback to large classes can become laborious. This could partially be automated for online assessments though. A good alternative would be to use rubrics (assessment grids), because they will tell the students exactly what was expected of them and on which level they performed. We will discuss how to go about this in detail further on in this reader.

1.7. Regulations and guidelines for assessment

Your assessment plan should be in line with the various regulations in place for your faculty. In this section you will find some basic information on which laws, regulations and policies might apply and where to find them. These are listed in hierarchical order:

1.7.A. 5.1 Law

The *Wet op het hoger onderwijs en wetenschappelijk onderzoek* (WHW; Law on Higher Education and Scientific Research, unfortunately only available in Dutch) is the law that determines how the universities in the Netherlands are organised. It also states that each programme should have a *teaching and examination regulations* (TER).

1.8. TER: Teaching and Examination Regulations and IR: Implementation Regulations

All regulations regarding admission, tracks, education, exams, etc. can be found In the TER (in Dutch: Onderwijs- en Examenregeling, OER).

Article 4 in the TER describes the **programme's exit qualifications**. The exit qualifications are the 'learning objectives' of the entire programme. The combination of learning objectives of individual courses should cover the exit qualifications of the programme. It is up to all lecturers at TU Delft to ensure that students meet all exit qualifications by the time they receive their BSc or MSc diploma. It is therefore important to take note of the following:

- Which exit qualifications (also called final attainment level, or intended learning outcome of a programme) should your course should contribute to;
- Whether there is a number of courses that contribute to an exit qualification;
- If yours is the only course contributing to a specific exit qualification.

This has implications for the level at which you need to assess specific learning objectives and the importance of the assessments in your course. Furthermore, it influences with which course coordinators and lecturers you will interact to align and fine tune your learning objectives and assessment plans.

For each subject that could be relevant to your assessment plan, the applicable section (§) and article number(s) (Art) are given for Bachelor and Master programmes. The numbers are based on the model TERs and actual numbers can vary slightly per programme. Here is also a link to all TERs, IRs and R&G of BoEs for all bachelor and master programmes at TU Delft.

Table 4.
Overview of assessment related subjects (first column) that are covered in the Teaching and Examination Regulations (TER, second column)

Teaching and Examination Regulations	
Obligation to participate in practical exercises	§3, Art 11.2 §5, Art 23
Number and times of examinations per year. Refers to the IR.	§5, Art 16 & Art 17
Validity duration of examinations (and sometimes of partial examinations)	§5, Art 22
Type of examinations (assessment method): refers to the appendix (IR).	§5, Art 16
Oral exam: number of students that is assessed at the same time, number of examiners, public nature of the exam, identity of the student	§5, Art 18
Announcement of grades (when, how and possibility for appeal against grade)	§4, Art 19
When students are allowed to inspect their assessed work, the questions/assignments and the criteria used for grading (answer models/rubrics) (and make a copy).	§4, Art 20
When and how a discussion of oral or written exams takes place	§4, Art 21

Take the time to make sure that your course assessments are in line with the requirements.

1.8.A. 5.3 Rules and Guidelines from the Examination Board/Board of Examiners

The 'Board of Examiners' (BoE) appoints the examiners to conduct examinations. Secondly, it checks the quality of the assessment of a programme. In addition, it grants exemptions to individual students and decides what measures will be taken in case of fraud.

In the 'Rules and Guidelines Board of Examiners' (R&G BoE), you can find a lot of information that is applicable to many stages of the assessment cycle (see page 19 of 'How to assess students through assignments' by Evelyn van de Veen, 2016), namely on test design, construction, administering and marking.

Table 5.
Overview of assessment related subjects (first column) that are covered in the Teaching and Examination Regulations (TER, second column), and Rules & Guidelines of the Board of Examiners (R&G BoE, third column)

Rules and Guidelines of the Board of Examiners	
Fraud	Art 7
Multiple examiners examining one examination	Art 10.1
(re)taking exams in different forms	Art 1.2-10.4
Online proctored examination	Art 11
Quality requirements of examinations	Art 12
Procedure during examinations	Art 13
Grading, rounding, partial grades, minimum grades, answer model	Art 14
Registering results in OSIRIS	Art 15
Archiving of work and results (duration)	Art 16
Projects	Art 20-21
Graduation projects	Art 22-25

1.8.B. 5.4 Assessment policies

At TU Delft, each faculty has developed their own *assessment policy* document that is based on the central assessment policy. The guidelines in these documents are usually very broad and general, but in some cases they contain very practical information that needs to be followed step-by-step. For example, it might contain regulations on when to exclude questions from calculating the final results, based on outcomes of the test analysis, or how to calculate a score (grade transformation). The assessment policies of the faculties and, if applicable, of programmes, can be found on intranet: https://intranet.tudelft.nl/en/-/assessment-policy-and-examination-guidelines

1.9. Checklist for quality requirements for assessment

In chapter 1 'Principles of assessment' in Van de Veen (2017), you will find a detailed description of the quality requirements for assessment.

The following table provides a checklist that you can use to evaluate whether your assessment plan as a whole, and your individual assessments meets the quality requirement of your assessment. Some of the requirements are explained in more detail here as well:

Checklist 1:
Summary of quality requirements for assessment

Quality requirement for assessment	Description
Validity	Validity is also called *representativeness* (whether the assessment represents the content and level of the learning objectives). This implies the following: - The tests cover the learning objectives and nothing else. - The tests are at the level of the learning objectives. - The assessment methods match the learning objectives. - The weighting of the LOs in the grade reflects the time spent on learning activities for each learning objective, as well as the importance of the learning objectives. Assessment *blueprints* (consistency checks tables for assignments and assessment matrices for exams) visualize whether an individual assessment represents the learning objectives.
Reliability	Reliability relates to consistency in grading and whether the student earns the grade that they are meant to earn. It can be split in test-retest reliability, specificity and objectivity: **Specificity** implies that: - Grades represent the level of mastery; - Only students who master the LOs to a desirable level can *pass* (for example: do not ask questions that students can answer on the basis of general knowledge or skills that are not specific to the course); - The *grade* should not be influenced by the assessment method. For example, the grade for a multiple choice exam mimics that of the open exam equivalent. - Measures to prevent *fraud, plagiarism, and free-riding* have been taken; **Test-retest reliability** implies that the same student should get the same score if they answer a question twice: - Questions should be **clear** enough for students to give the same answer 5 minutes later (and therefore get the same amount of points); - Exams should have the same *difficulty* over the years; - *Enough questions* are asked to get a good sample. **Objectivity** implies that the grade does not depend on the grader (rater), i.e. the *rater bias* is minimised.
Transparency	Making grading criteria and methods known and clear to students: - Before the assessment (preparation required, example questions, weighting of learning objectives); - During the assessment (points per item/criterion, cut-off score/grade calculation); - After the assessment (calculation of grades, feedback on errors).
Practicability	Also referred to as ‘usability’. This relates to the workload and availability of resources, for example: - It should be possible for students who do well to get a 10, within the hours stipulated for the amount of EC that they have to work; - How feasible is it for the lecturer(s) and teaching assistants to prepare, provide feedback and grade the assessments?

Quality requirement for assessment	Description
Efficacy	Efficacy is the extent to which the assessment plan and the individual assessments **stimulate student learning and mastery of the learning objectives**. The following questions may help you: - Is the assessment *authentic* (i.e. is it comparable with what the student will be doing in the real world of work)? - Does the assessment stimulate learning? - Is the *feedback* effective for the student? ○ Do students get *feedback* on their performance on each learning objective *before* taking a summative assessment? ○ Is the feedback focussed on learning objectives? ○ Do the students get the feedback in time to improve their performance before their next assessment? ○ Is the feedback specific enough (by focussing the feedback on the criteria and informing the students what the next step is to improve on a criterion)? Is the assessment effective in such a way that you as a lecturer can adapt the course on the fly (for example, by giving extra exercises or omit learning activities)?

Using the quality requirements to improve your assessments can improve the quality of your course as a whole. You might find, however, that optimising your assessment for one of the requirements compromises the level of quality according to another requirement. There will almost always be a trade-off, so it is up to you to decide what is most important for your students and your course.

For example, medical students might not always get the opportunity to perform certain procedures on real patients during their studies. However, they still have to be evaluated. Mock-ups are usually used to simulate scenarios (making the assessment practically feasible), but this compromises validity of the assessment.

The grades you assign to your students can have far-reaching consequences for the continuation of their studies, scholarships and perhaps even on their careers. For that reason, it is important to know what to do when students obtained low grades because of an issue in the learning activities, assessment or grading process. In this section, we will discuss grade calculation, and alterations that could be made after a test result analysis.

1.10. What is a grade?

The meaning of a grade is described in the Rules and Guidelines of the Board of Examiners of your programme. In general, it looks like this (R&G BoE master's programmes MSc AP/CE/ LST/NB/SEC):

9,5 – 10,0	Excellent
8,5 – 9,0	Very good
7,5 – 8,0	Good
6,5 – 7,0	More than satisfactory
6,0	Satisfactory
4,5 – 5,5	Unsatisfactory
3,5 – 4,0	Poor
1,0– 3,0	Very poor

More importantly, the grade should relate to how well a student masters the learning objectives. If students demonstrate in a test that they master all learning objectives, they should be awarded a 10. A 1, on the other hand, is by Dutch convention the lowest grade that a student can obtain.

1.10.A. What does a minimum pass grade imply?

A 6 (or 5.8 before rounding) is *the minimum pass grade*. It implies that a student (on average) masters the learning objective at the *minimum level* to a) pass this course, and b) either start a course that builds upon this one, or in case there are none, c) master the related final attainments of the bachelor or master programme at the minimum required level and start their professional lives.

The responsible lecturer should determine what the *minimum level* at which the students will get this *minimum pass grade* (6.0). If a course is assessed with an exam with open questions, students often get a 6.0 if they receive 60% of the maximum score. Higher or lower percentages are also possible. Depending on the level of the questions, this may imply that a score of 6.0 implies that a student on average masters 60% of the learning objectives. They may not master some LOs at all, and may fully master other LOs. The exam averages this out.

For a master thesis that is assessed using an assessment sheet with scores on different criteria, it may imply that a student at least masters *each individual criterion* up to 50% (otherwise they would not have gotten their green light meeting), and that on average they master the criteria (that should be aligned with the learning objectives) at the minimum levels that are described in the assessment sheet. As you can see, in an assessment sheet for assignments and projects, it is possible to require a minimum level for certain criteria / learning objectives.

1.10.B. Introducing more tests requirements to test LO achievement?

What about having one exam per learning objective, and requiring a 5.0 for each and every one of them? Or adding minimum levels for each criterion in assessment sheets of assignments and projects? Increasing the number of assessments?

There is a large objection against increasing the number of assessments. We are unable to create perfect assessments that perfectly measures the 'true' extend to which a student masters the learning objectives. The resulting 'measurement error' can be as high as two points on a grade from 1-10. If we increase the number of tests and their accompanying minimum grades, we increase the number of students who will fail the course incorrectly. Keep in mind that in the Rules & Guidelines of Examiners, in article 14, partial grades often require a minimum grade of 5.0.

Furthermore, students need to have resits for all these extra hurdles, which would mean a lot of work for the lecturer. Furthermore, studying for resits or working on additions will steal away time from the other courses in the next period and therefore deteriorate student performance in the next period. Therefore, we should be careful not to create unnecessary assessment 'hurdles'.

On the other hand, it is important to have insight on which learning objectives are accomplished by the students, and which not. Your course is, after all, part of a larger program and qualification. Once a student graduates, it is assumed that the students have met the outcomes.

To conclude, be aware that introducing extra assessments will come with extra resits and additions, and therefore extra work for students and teachers. Carefully balance the need to ascertain a minimum level for important learning objectives in the light of being able to successfully take follow-up courses and reaching the final attainments, with practicality for teachers (extra reviewing) and students (studyability of the next period in which they have to repair deficiencies).

1.10.C. What does a pass mean for the follow-up course?

Another question that is important to consider is the following: What guarantee does a pass give to the student about success in the rest of his study, and what guarantee does a pass give to your colleague that the student is able to successfully follow his or her course (this is called criterion validity)?

Let us take a course on Electricity as an **example**. The course coordinator (also called the *responsible lecturer*) assumes the students have acquired the necessary mathematical skills to solve the equations, since it was a learning objectives of the previous course. What can the course coordinator expect of her students on this 'achieved' mathematical learning objective? What if the students learnt this learning objective at the level of a 6? And what if the students skipped this learning objective in the last course and still managed to pass the exam?

It might be a good idea to talk to course coordinators of preceding and succeeding courses to discuss and (re)define the desirable level of a 6, so that they know what level they may expect from the students. It is unrealistic to assume that students master a learning objective of a previous course at the level of the learning objectives (a 10). Talking to colleagues will also enable you to give students advice on where to find information and (extra) exercises without you having to design the exercises and other material yourself.

1.11. Grade calculation

1.11.A. Score-grade transformation and cut-off score for open-ended questions

After grading an exam or assignment, you usually end up with a *score*, which is a number of *points*. Now, you have to decide on the grade that corresponds to the points, that is, you do a *score-grade transformation.*

Possible formulas for score-grade transformation in open questions (graphical representation in Figure 2):

Light blue squares: $Grade = 1 + 9 * (\frac{score}{\max(score)})$

Dark blue triangles: $Grade = \max\{1; 10 * \left(\frac{score}{\max(score)}\right)\}$

With *Grade* the calculated grade, *score* the obtained score by the student, *max(score)* the maximum obtainable score for the assessment, and *max{a;b}* the maximum value of a and b.

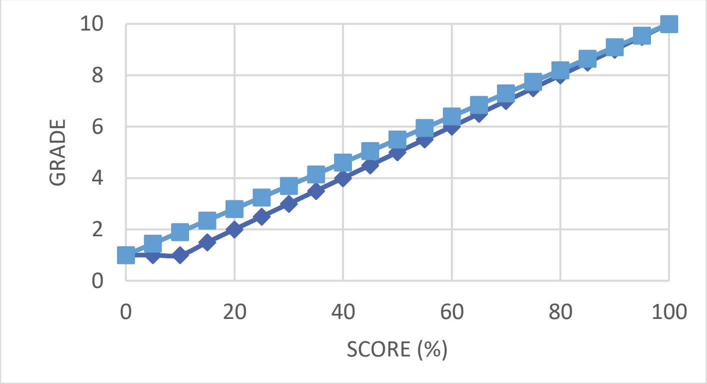


Figure 2. Two simple score-grade transformation. Horizontal axis: relative score (percentage). Vertical axis: grade.
Light blue squares: 0 points lead to a 1, the grade increases after each point earned.
Dark blue triangles: grade runs from 0 to 10 and rounded to 1 for grades smaller than 1.

The method you choose will determine on the *cut-off score*: the *cut-off score* is the number of points a student needs to obtain in the exam in order to obtain the *minimum pass grade*.

In Figure 2, for the dark blue triangles, a 6 corresponds to collecting 60% of the points, while for the light blue squares, 55% of the points will assign the student a 6.

Remember to communicate the *cut-off score* to your students on the exam's cover page or assignment instructions.

1.11.B. Score-grade transformation and cut-off score for closed-ended questions

When calculating the grade for MCQs, you are advised to adjust the grade to compensate for guessing. This is called 'guessing correction'. Statistically speaking, students who are unfamiliar with the course content can score a percentage of correct answers that is inversely related to the number of answer options.

The reason for applying a correction for guessing can be found in quality requirement *reliability*, which implies that the question type (open, closed, etc.) should not influence the grade. If students do not know anything about the course content, they should get a grade of 1.0, regardless of whether the exam had open-ended or closed-ended questions.

For example: in case of 4 options (1 correct answer and 3 distractors), the guess correction is ¼ = 25%, and for true/false questions, the guess correction should be 50%. For an exam with 54 questions, with 3 options each the guessing correction is 33.3% * 54 = 18 points. Grade = 1 + 9 * (points – guessing correction)/(54 – guessing correction) = 1 + 9 * (points – 18)/36.

If it were an open question exam, they would get 0 points.

Because you want your students to get the same grade for an MCQ-test as for a test with open-ended questions (for *reliability*), you would subtract the number of points they can earn by guessing, from the total score. In the score-grade transformation of multiple choice questions, the guess correction should be taken into account, such that the students will have no points (or a 1) whenever there score is equal or lower than the guessing correction.

Possible linear formulas for score-grade transformation for closed-ended questions are:

$$Grade = \max\{1; 1 + 9 * \left(\frac{s-gs}{(\mathrm{ms}-\mathrm{gs})}\right)\}$$

$$Grade = \max\{1; 10 * \left(\frac{s-gs}{(\mathrm{ms}-\mathrm{gs})}\right)\}$$

with *Grade* the resulting grade, *s* the obtained score by an individual student, *gs* the guessing score (average obtained score of random guessing), *ms* the maximum score, and max{a;b} the maximum value of a and b.

1.11.C. Setting the cut-off score manually & resulting score-grade transformations

The previous grade calculations automatically resulted in a cut-off score.

You can also decide on an appropriate *cut-off score* yourself. You determine this score by determining for each subquestion how many points a student with a 6 would on average gain for this question. The sum of this is the cut-off score. **The cut-off score should reflect the minimum level that students should have reached in order to pass the course.** This then should pave the way for students to pass follow-up courses, and achieve the exit qualifications of the programme to an acceptable level.

In the next paragraph, you read how to set the cut-off score manually.

If you want to set the cut-off manually, you will need to split the score-grade transformation around the cut-off score. In Figure 3, you can find a graphical representation of traditional (green), and split score-grade transformations with a cut-off score of 16 points (grey x) and 32 points (blue filled squares) respectively. This representation is for closed-questions.

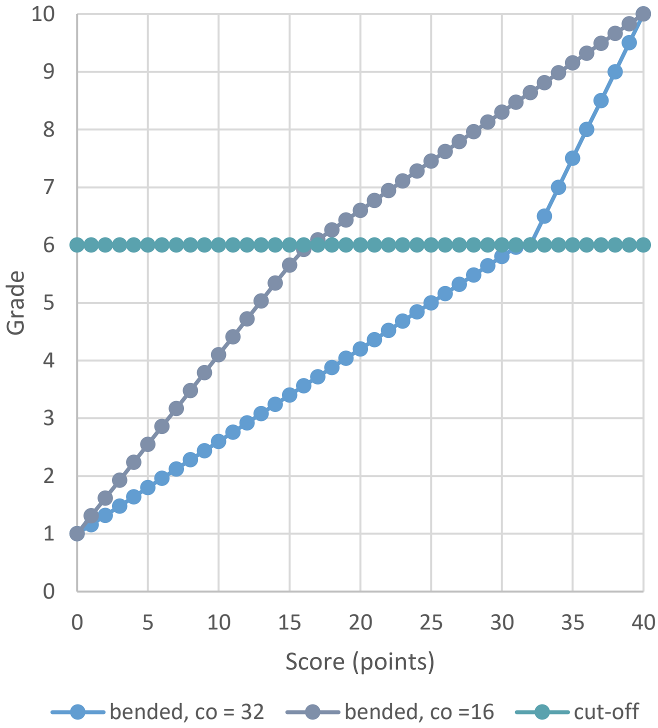


Figure 3. Score-grade transformations for two split transformations around cut-off scores of 16 points and 32 points. cos = cut-off score. Formulas see running text.

Used formulas:

$$Grade = \begin{cases} 1 + s\dfrac{5}{cos}, & s < cos \\ \dfrac{6ms - 10cos + 4s}{ms - cos}, & s \geq cos \end{cases}$$

with *Grade* the resulting grade, *s* the obtained score by an individual student, *gs* the guessing score (average obtained score of random guessing), *cos* the cut-off score, *ms* the maximum score, and max{a;b} the maximum value of a and b.

1.11.D. Closed-ended questions, knowledge percentage and relation to cut-off score

In exams, the *knowledge percentage* is the percentage of questions that a students should be able to correctly answer to reach the *cut-off score and minimum pass-grade*.

For multiple choice questions the *cut-off score* is higher than the *knowledge percentage* times total number of questions.

For example, if you want your students to answer at least 60% correctly of an open-ended question (i.e. *knowledge percentage* is 60%), your *cut-off score* in case of MCQs with 4 options needs to be 25% (*guessing score*) + 60% (*knowledge percentage*) x 75% (100% - guessing score = remaining score) = 25% + 45% = 70%. **In other words, students will get a pass when they correctly answer 70% of the questions (*cut-off score*), for a *knowledge percentage* of 60%.**

Consider the following questions:

- At pass level, what knowledge level (%) do students have?
- Is a knowledge percentage of 60% too low and should the students meet more criteria per learning objective to deserve a pass?

If the learning objective is to design a bridge, is it enough if the students meet 60% of the design specifications, or is it important that *all* of them are met? What impact could it have on their careers if they only meet 60% of the requirements? How much will they use what they have learned in your course? What year are the students in? Will there still be a course that builds on this learning objective or is your course the last one in the programme where your students should perform on the level of the exit qualifications of the programme? What are these exit qualifications (you can find them in your programme's TER)?

1.11.E. How question difficulty influences cut-off score

How difficult should an exam or assignment question be? It depends on whom you are asking. From an item analysis point of view, it is best if the average score is low, for example 50% (i.e. with a p-value of 0.5). However, it might make students and lecturers feel demotivated when on average only half of the questions were answered correctly. Furthermore, students should demonstrate what they *are able to do* during an exam, and not what they *are not able to do*.

For both the students and lecturer, it is important to distinguish between a 6, 7, 8, 9 and 10. These grades could give a good indication of the student's level of achievement. On the other hand, a 1, 2, 3, 4 and 5 all result in a 'fail', regardless of the grade. If 58% of your points would lead to a 5.8 (pass), that would mean that you have only 42% of the exam points left to distinguish between the range of 6-10.

Let us say that your exam has 40 points to divide in steps of 1 points, that would mean that a change of 1 point changes the grade by 0.23 ((10-1)/40), so the step-size of one point is 0.225 grade. If you would have 60% of the points left (i.e. a cut-off at 40% of the points), the step-size will be smaller for the pass grades, i.e. 0.17 points ((10-6)/(60%*40)), and coarser for the fail grades, i.e. .31 ((6-1/(40%*40))).

If you choose a lower cut-off score, you have more points left to distinguish between the grades of 6, 7, 8, 9, and 10. 50% of the points could imply a 7.0, for example. One way to do this is to determine the number of points at which a student will have a 6.0 (the cut-off score). You can then linearly interpolate between 0 points (1) and the cut-off score (e.g. 15.0 points, 6), and between the cut-off score (15.0 points, 6) and the maximum score (40.0 points, 10). This is demonstrated in **Error! Reference source not found.** with the grey x-symbols. The gradient of the line changes at the cut-off score (arrow): the line is shallower between the cut-off score and the maximum score.

If this exam would be very difficult but would still result in a high pass rate, due to the low cut-off score, this could imply that students would pass while they could answer only very little questions. This may demotivate students quite a lot (and may demotivate you too, while grading). Furthermore, constructive alignment and transparency demands that your students practice with questions that are at the same level as the exam. You and your students would be worried if they would only be able to answer 50% of the questions after having completed your course.

> To conclude, theoretically, students should on average score 50% (p = 0.5) on all questions, and you can choose a cut-off score below 50%. However, aiming for an average score of 50% might leave both students and graders depressed. Find a mix of both challenging and few easy questions, that will help you to distinguish grades between 5.0 and 10.0. Make sure that the easy questions cannot be answered without actively participating in your course.

1.11.F. Exams with both open-ended and closed-ended questions

If your exam consists has open-ended *and* closed-ended questions, you are recommended to calculate a grade for the open-ended questions, and a grade for the closed-ended questions *separately*. Then, for the grade calculation of the closed-ended questions, also, you must take into account the guessing correction. After calculating both grades, you average calculate the total *weighted average*. Communicate the weighting of both grades to your students (before, during and after the exam). It is helpful for students to know the separated grades, too, since it gives shows them feedback on what type of questions they need to focus most during their preparation for future assessments.

The *reason* why you need to calculate the grades separately, is so that the guessing correction must be done on the points of the closed-ended questions. The following example will illustrate *why* you are advised to calculate the two grades separately.

Let's assume that the exam consists of 100 points:

- 60 points to be earned in open questions

- 40 points divided over 40 MCQs with four alternatives

- In order to correct for guessing, 10 points need to be deducted from the score.

Now let's assume that one of our students did not get any points for the MCQs (0 points) and full points for the open questions (60 points).

Firstly, we consider the situation in which we apply guessing correction, and calculate a combined grade at once. Because of the guessing correction, the corrected amount of points would be 50 (60 – 10 points), out of the maximum of 90 points (100 – 10 points). Depending on the calculation, this would lead to the student attaining a 6.0 or a 5.6 (see).

Secondly, if we apply guessing correction and calculate separate grades, the grade varies between 6.0 and 7.0, depending on the ratio of the weights of the open question grade and closed-ended questions. The technical reason for the difference is that in case of combining the grades, the grade for the closed-ended questions is virtually negative (see).

However, in order for the grade to represent the level of learning objective achievement, it is undesirable to have negative grades, especially since the grading of closed-ended questions should be comparable to the grading of open questions. For an open (sub)question in an exams, you would not give negative points when a student would not fill in anything for a certain subquestion, nor when he would have made an enormous amount of errors within this subquestion. The minimum amount of points per subquestion is 0.

Concluding: In order to prevent (virtual) negative grades (or points) in case of guessing correction, you are advised to use the weighted average of the MCQ grade and open question grade.

Table 6.
The influence of grade calculation decisions on grades for exams with a combination of open and closed-ended questions, for three hypothetical students with different scores for both question types. Ratio open scores vs MCQs: 60:40

		Grading student A open questions: 60/60 MCQs: 0/40			**Grading student B** open questions: 60/60 MCQs: 10/40			**Grading student C** open questions: 30/60 MCQs: 20/40		
	Separate grades or single grade?	**Grade open questions**	**Grade closed questions**	**Total grade**	**Grade open questions**	**Grade closed questions**	**Total grade**	**Grade open questions**	**Grade closed questions**	**Total grade**
Guessing correction	**Separate grades**	10,0	1,0	6,4	10,0	1,0	6,4	5,5	4,0	4,9
	Single grade	10,0	-2,0	6,0	10,0	1,0	7,0	5,5	4,0	5,0

In the table, you will find the difference in grading for including or excluding guessing correction (first column), calculating separate grades for open and closed questions or not (second column) and if so, the ratio between the open and closed questions (third column), and whether the increase of (sub)grades start at 0.0 or 1.0 (fourth column). The results are displayed for three students: student A has full points for the open questions and no points for the closed questions, student B has full points for the open questions and guessing score for the closed questions, and student C obtained half points for both open and closed questions.

1.12. The process of grading: Increasing objectivity and reliability, while decreasing grading errors

In this section, objectivity, or the reliability of the grade is discussed, as well as possible solutions for errors made by assessors. Because we are human, it is nearly impossible for us not to occasionally make errors when grading. There is also even more room for errors when more than one assessor is grading the same assessment - different assessors will simply grade differently. When assessing your students, it is important to at least be aware of this, and to take certain measures to prevent inconsistencies.

1.12.A. Inter-assessor reliability

It often happens that student grades partially depend on *which* assessor graded the work. This is mainly because the following happens during grading:

- ***GENEROSITY ERROR:*** assessors are (too) lenient;
- ***SEVERITY ERRORS:*** assessors are (too) strict.

To help prevent, or at least diminish these errors, it is recommended to follow these guidelines:

- Use a detailed answer model or rubric. This leaves less room for assessors' own interpretation.
- Use two assessors per sample of students' work to even out differences in interpretation.
- Distribute the questions - not the students - over the assessors. This way all students are evaluated equally generous/strict.
-
- Have a session in which all assessors discuss the meaning of the answer model. Then grade a few samples of students' work and discuss and resolve any differences in rating. Only when everyone seems to interpret the results consistently, the actual grading can begin.

1.12.B. Intra-assessor reliability

When there is just one assessor who evaluates all students' work, there are a number of factors that endanger objective and reliable evaluation of students' results. Here are three examples:

HALO AND HORN EFFECT: the assessor allows their general impression of the student influence the scores.

- Mark the test anonymously by having students only write their student number on the answer sheets.
- Let someone who does not know the students evaluate the results.
- Have two assessors - one of which does not know the students - evaluate the results.
- Use an answer model or a rubric.

CONTRAST EFFECT: over- or underrating students' work because of the quality of other students' answers that were graded previously.

- Use an answer model or a rubric.
- Evaluate per question – not per student, and change the order of the students per question.
- Rescore the first few samples after you have finished all. The first ones are usually scored more strictly then the rest.

- ***SEQUENCE EFFECT:*** shift in standards, or redefining the scoring criteria over time.

- Use an answer model or a rubric.
- Evaluate per question – not per student, and change the order of the students per question.

When considering how well we and our students performed, we are frequently asked to report the percentage of the students that passed the course. However, analysing their scores will reveal more detail and enable you to make informative decisions for improving the assessment and the course as a whole.

A test result analysis will give insight into:

- How well the students mastered the individual learning objectives of the course;
- The overall quality of the assessment;
- The quality of the individual test questions or assignment criteria; and
- Whether the answer model and/or the grading would need to be revised.

In this chapter we will explain the steps that you can take to perform a test result analysis and to improve the grading, future assessments and future courses based on the findings.

Before we start, please take note of the following: By 'test', we mean any assessment, including projects, assignments, exams with open-ended questions and multiple choice exams. By 'grade', we mean the grade (usually on a scale from 1 to 10) that a student receives for the whole test, and by 'score', we mean the number of points that a student obtained from this test, before it is transformed into a grade. An 'item' is the smallest unit in a test. This can be a criterion or subcriterion for assignments/projects, or a subquestion or question for an exam.

If you are reading through this chapter as part of UTQ module ASSESS, the purple/blue bar in the left margin of the text indicates that this text is part of UTQ module ASSESS. The rest is background information, and not necessary to read.

1.13. Format of test result data

First of all, to conduct the analysis, you will need data in a spreadsheet that you collect once your students have completed an assessment. You will use the following:

- Scores per item for each student;
- Total scores per student.
- In case the test grades are not calculated directly from the points per, you also need the grades per student.

1.13.A. Exams

If you are analysing an exam, the format of your spreadsheet would look more or less in

Table 7. You will need the points per student, per subquestion.

Table 7.
Example of exam data

maximum points	0,5	1	2	1	0,5	1	0,5	2
studentID	1a	1b	1c	1d	2a	2b	2c	2d
123456	0,5	1	1	1	0,1	0,2	0,0	1
123457	0,2	0	2	0,7	0,5	1	0,5	2
123458	0,5	1	1	1	0,4	0,8	0,4	1
123459	0,4	0	2	0,9	0,3	0,6	0,3	2
123460	0,4	0	0	0,9	0,1	0,2	0,0	0
123461	0,4	0,5	1	0,9	0,2	0,4	0,1	2
123462	0,4	1	2	0,9	0,2	0,4	0,1	1
123463	0,5	0	1	1	0,2	0,4	0,1	0
123464	0,2	0,5	0	0,7	0,5	1	0,5	2
123465	0,5	1	2	1	0,4	0,8	0,4	2
123466	0,2	0,5	1	0,7	0,4	0,8	0,4	1
123467	0,5	0	2	1	0,2	0,4	0,1	1

1.13.B. Multiple choice exams

In case of MCQs, you need to have the data with the answers of the students, as shown in Table 8.

Table 8.
Example of MCQ data

Possible answers:	ABCDE	ABCDE	ABCDE	ABCDE	ABCDE	ABCDE	ABCDE	ABCDE
Score if correct:	5	5	8	8	8	8	4	6
Correct answer:	C	E	B	A	E	C	A	A
	1	2	3	4	5	6	7	8
	C	E	B	D	A	C	B	A
	C	E	A	A	C	C	A	A
	C	E	A	A	D	C	B	A
	C	E	B	A	E	C	A	A
	C	E	A	A	E	C	A	A
	C	E	B	D		C	B	A
	E	E	B	E	E	A	E	C
	D	E	E	A	E	C	E	D
	C	E	D	A	D	D	A	D
	C	B	E	A	D	C	A	E
	E	E	E	E	E	C	B	A
	B	E	A	A	E	C	A	A
	C	E	D	E	E	C	B	A
	D	E	B	E	B	D	C	B
	C	B	A	E	E	C	E	E
	D	E	A	B	B	B	B	A
	C	E	B	A	E	C	A	A
	E	E	A	A	E	C	B	B
	C	E	A	D	D	C	A	A
	C	E	B	E	E	C	B	A
	C	E	B	A	E	C	B	A
	C	F	A	E	E	C	B	A
	C	E	B	A	E	C	A	A

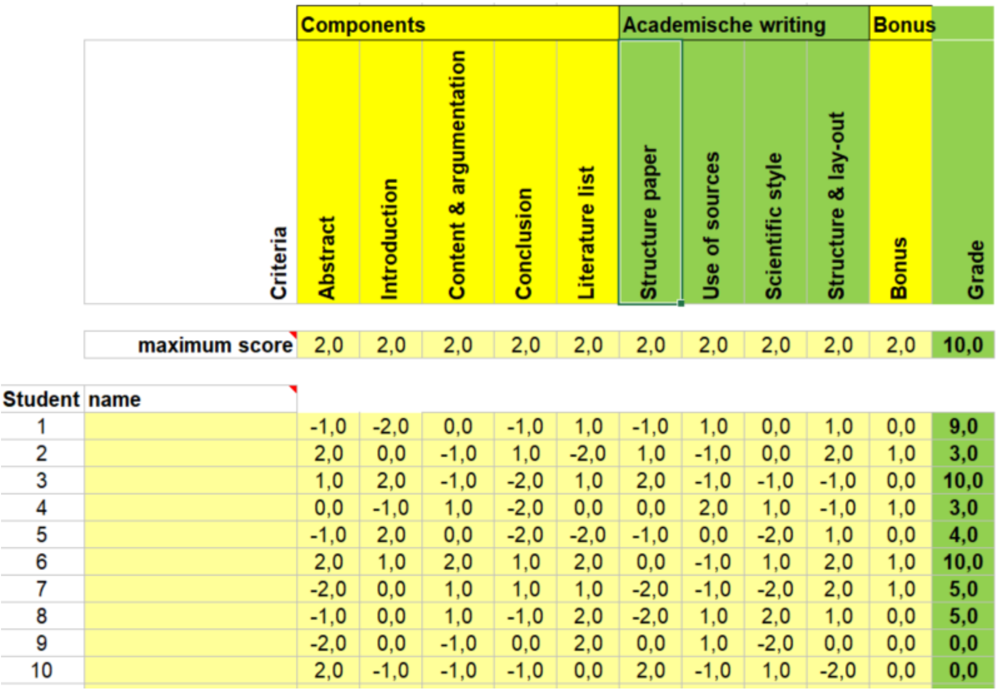
1.13.C. Assignments or projects

In case of an assignment, the data should look more or less like in

Table 9. You will need the points per criterion.

In the example figure, you can see that the criteria are grouped ('components', 'academic writing' and 'bonus'). This grouping is not mandatory. Since we consider it to be very important to analyse assignment and project data too, and since little tooling is available, we are happy to extend the available Excel sheet to your needs. Please let us know if you have for example more criteria, or if you would like to compare the scores of different graders. We will adapt the Excel sheet to your needs. In the future, we want to use your requests to improve the standard Excel sheet for assignments or projects.

Table 9.
Example of assignment data



		Components					Academische writing				Bonus	
	Criteria	Abstract	Introduction	Content & argumentation	Conclusion	Literature list	Structure paper	Use of sources	Scientific style	Structure & lay-out	Bonus	Grade
	maximum score	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	10,0
Student	name											
1		-1,0	-2,0	0,0	-1,0	1,0	-1,0	1,0	0,0	1,0	0,0	9,0
2		2,0	0,0	-1,0	1,0	-2,0	1,0	-1,0	0,0	2,0	1,0	3,0
3		1,0	2,0	-1,0	-2,0	1,0	2,0	-1,0	-1,0	-1,0	0,0	10,0
4		0,0	-1,0	1,0	-2,0	0,0	0,0	2,0	1,0	-1,0	1,0	3,0
5		-1,0	2,0	0,0	-2,0	-2,0	-1,0	0,0	-2,0	1,0	0,0	4,0
6		2,0	1,0	2,0	1,0	2,0	0,0	-1,0	1,0	2,0	1,0	10,0
7		-2,0	0,0	1,0	1,0	1,0	-2,0	-1,0	-2,0	2,0	1,0	5,0
8		-1,0	0,0	1,0	-1,0	2,0	-2,0	1,0	2,0	1,0	0,0	5,0
9		-2,0	0,0	-1,0	0,0	2,0	0,0	1,0	-2,0	0,0	0,0	0,0
10		2,0	-1,0	-1,0	-1,0	0,0	2,0	-1,0	1,0	-2,0	0,0	0,0

1.13.D. preparation of data (all test types)

Each column contains an 'item'. An item is either a subquestion in an exam/exam-like assignment, or a criterion that is used when grading an assignment/ essay-like exam question.

For the analysis, it is important that the cells contain points that do not need to be weighted. For example, if you give your students a grade (1-10) for each criteria, and afterwards apply a weighing to each criterion to calculate the grade, you will need to multiply the grades with the weighing criteria before copying it to the table.

For example, student A has a 7.0 for 'code style', which has a weight of 20% of the grade. This student will get a 7.0*20% = 140 points. The maximum points is 10*20% = 200 points. You could also use 1.4 points and a maximum of 2.0 points, respectively, as long as you treat all items the same.

The test result analysis works best if you have enough items (subquestions in exams or criteria in larger assignments) that determine a final/partial grade. This is usually the case for the following tests:

- Exam with open-ended questions with at least 10 (sub)questions. Make sure to put the results of each subquestion in a sheet.
- Exam with closed-ended questions (for example multiple choice questions or true/false questions) with at least 160 answer possibilities (for example 40 multiple choice questions with 4 options). You need data that includes the chosen answers (e.g. 1A, 2C, 3A for a certain student).
- Assignment that has at least 10 scored items. For example, a programming project in which groups of students need to deliver a project (scored on 5 criteria) and report (scored on 3 criteria), and are individually scored on programming skills and academic attitude (2 criteria).
- A set of at least 8 assignments/other tests that determine the final score.

For smaller datasets, you may not be able to draw strong conclusions from your data. However, you are encouraged to use your own data, whenever possible. This will make the analysis as useful as possible for you. Please contact Lisette Harting to discuss options of analysing and interpreting your data.

1.14. Available tools

- **Excel**: you can use an Excel sheet[2] for open and closed questions, or the Excel sheet for assignments and projects. Make sure to fill in the data per subquestion (i.e. consider 1a, 1b, 1c are individual questions) or assessed criterion (in case of assignments and projects).
- **Contest**: in case you use paper-based multiple choice questions that are processed automatically at the Education and Student Affairs (ESA), you can analyse the resulting data in Contest. You need an account to login. More info here.
- **SPSS**: you can use the Technical University of Eindhoven's manual to use SPSS to generate results (in English!)
- **Matlab**: there is a Matlab-script available that transforms Excel-data to boxplots, *and* performs a test result analysis.

1.15. Do students master the individual learning objectives?

The first question you want to ask yourself, is how well your group of students master the individual learning objectives. Are they performing better at certain learning objectives than others? Did my new teaching approach for a certain learning objective work? Are there learning objectives in which they perform worse than in others? These and other questions might be answered by grouping the (normalized) item scores per learning objective, like in Figure 4.

Graphically summarizing the scores of your students per learning objective will make this easier to interpret the results. Plot a measure of performance (average and/or median) and spread (standard deviation or boxplot), and if helpful, the individual datapoints.

When analysing the graph, think about what scores you as a teaching professional find acceptable for a particular course or learning objective. Also consider what caused the problems or success in LOs during the course, and how you can help your colleagues and students to work on (and prevent) knowledge gaps.

Typically, problems in learning objective achievement are caused by a lack of practice at test level (constructive alignment). You can use any graph of your choice, as long as it summarises the distribution of the scores per learning objective.

[2] Available in Brightspace

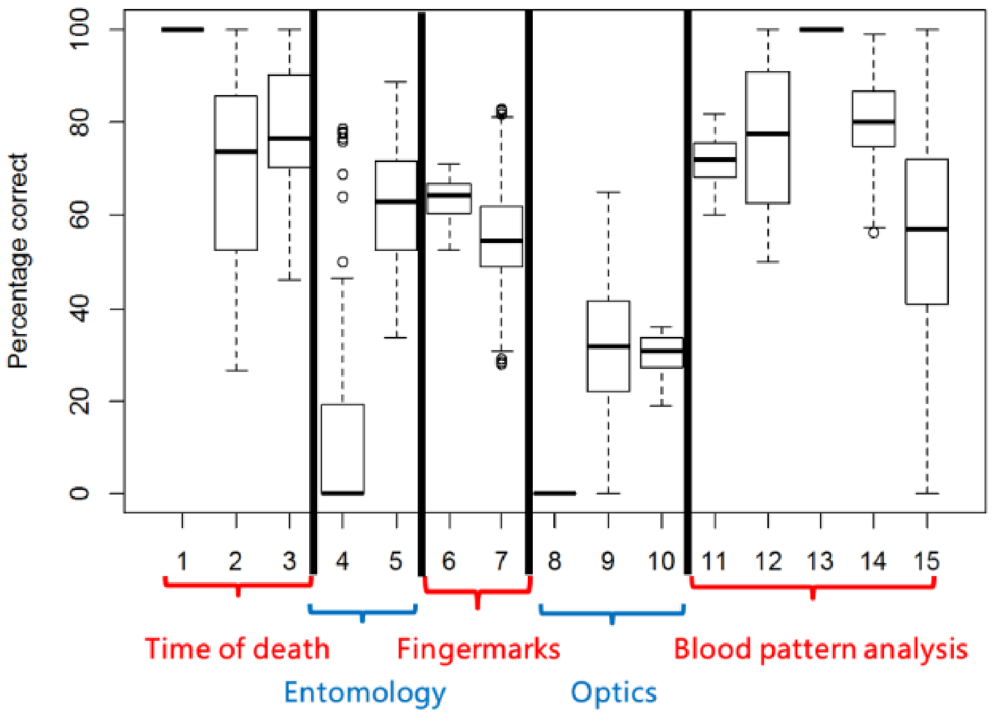


Figure 4. Example graph indicating test scores and LOs in a boxplot

Here are some more tips:

- You can use the relative scores (percentage or fraction) instead of absolute number of points if you want to compare performances between items.
- If you want to use boxplots (not for multiple choice questions or other closed-ended questions with only full or no points!), Excel is not the only option. R and Matlab have a better box-plotting functionality. There is a Matlab-script available that will help you plot the boxplots. In the available Excel sheet, you will find a link to an online box-plotting tool, to save you time. Boxplots is preferable to plots that use standard deviations as measure for the spread, in case your data is not normally distributed, as is often the case for grades.
- If you are using MCQs where students can only get no or full points per question, boxplots are not of any use, since students will either have 0 or 1 point (in general). You can plot the average score which corresponds to the percentage of students who got a question correct.

1.16. Reliability of the test (Cronbach's alpha)

Please watch the **video** on the difference between reliability and validity in test result analyses.

The reliability of a test is the same as the reliability of the grade. Does the student with a 6.0 really deserve to pass, or are we not so sure, due to measurement errors? One way to estimate the measurement error is to calculate the score reliability (reliability coefficient), like Cronbach's alpha.

All reliability coefficients assume that the test intends to measure one single thing, namely how well as student masters a course. It also assumes that each student should perform more or less equally well on all test items, considering the fact that our job as teachers is to help student master *all* learning objectives of a course. If our students participate in all learning activities of our constructively aligned courses, it would be highly unexpected and worrisome if the highest performing students would have the lowest scores on the easiest questions, or the other way around.

Reliability coefficients are a measure of whether students are performing consistently well on all test items. This is also called the *internal consistency of the test*. It is the extent to which the outcome of the assessment is *not* influenced by coincidence.

There are several methods for calculating the reliability of an assessment. Cronbach's alpha is one of these methods. It estimates the test-retest by considering each question in the test as a separate test and then calculating the correlation between the questions. A simplified version for multiple-choice exams is KR-20.

The **value of** α always lies between 1 and 0. The closer the value is to 1, the smaller the measurement error. A lower reliability can mean that a student whose 'true score' is just above the cut-off score may fail the test due to test inaccuracy. Test reliability is very important when the consequences of the test results are large, and therefore the reliability coefficient should be higher for tests of higher stakes.

Grades can be considered reliable if Cronbach's alpha is high enough. This depends on the importance of the assessment (van Berkel, 1999):

Type of assessment	Cronbach's alpha
High stake assessment (e.g. only assessment of course	α ≥ 0.8
Medium/low stake assessment (e.g. 50% of final grade):	α ≥ 0.7
Formative assessment (e.g. 0% of final grade)	α ≥ 0.6

If your reliability is low, this may be due to the following factors (van Berkel, 1999):

- Test length: There may not be enough items in the test, which diminishes the reliability.
- Group composition: a more heterogeneous group of students leads to lower reliability, since some students might be good at e.g. the math part of the test, and other students might perform better at other questions. This can be an indication that you might want to tailor your course for these two groups and have your students practice on their weak points. This will increase Cronbach's alpha, as well as the item correlations (see 1.19.C on page 36).
- Test heterogeneity: If the items represent very different topics or skills, this will lead to a lower reliability coefficient.
- Difference between average item score (difficulty): the reliability coefficient will be lower if there are little items of average difficulty, and mostly items that result either in a low score in most students, or a high score in most students.
- Difference between student levels: the reliability coefficient will be lower if students are at more or less the same level.
- Item reliability: lower quality items (with higher Rir) decrease reliability of the entire test (see 1.19 to analyse this in detail).

The formula for calculating the reliability coefficient Cronbach's alpha is as follows:

$$\alpha = \frac{K}{K-1} \cdot \frac{\sigma_x^2 - \sum_{j=1}^{K} \sigma_j^2}{\sigma_x^2}$$

With α the reliability coefficient, K the total number of subquestions, σ_x^2 the variance in the final scores of all students, i.e.:

$$\sigma_x^2 = \frac{1}{N_{stud}} \sum_{i=1}^{N_{stud}} \left(x_i - \mu\right)^2$$

With N_{stud} the total number of students, x_i the final score of student i, and μ the mean final score.

The variance of the subquestion scores σ_j^2 is calculated equivalently:

$$\sigma_j^2 = \frac{1}{N_{stud}} \sum_{i=1}^{N_{stud}} \left(s_i - \mu_j\right)^2$$

With s_i the subquestion score of student i on subquestion j, and μ_j the mean score on subquestion j.

The reliability coefficient gives an indication of the reliability of the test as a whole by comparing the difference of the variance in the final test scores of all students with the variance in the test score per subquestion. The reliability coefficient can have a value between 0 (unreliable) and 1 (reliable). In very rare cases it can be negative. With a reliable test, the variance in the final scores of the students should be much larger than the sum of variances in the subquestion scores.

1.17. Confidence interval of grades (SEM)

To illustrate the meaning of reliability, we will now discuss how you can use Cronbach's alpha to calculate the measurement error that was introduced by chance. Just like any other measurement instrument, an assessment can also have a measurement error. We will first discuss the standard error of measurement (SEM) or the 68% confidence interval (in points), which we will transform to a confidence interval in grades.

Test theory assumes that *every student has a true score*, which reflects that student's actual capability in the area of expertise that an assessment is testing. If a student would take the same test an infinite amount of times, the average of all these scores would constitute the true score. Because this would not be practical to carry out, it is important to recognise that *the score of a student taking a test once is always the measurement of the true score plus the measurement error,* either systematic of accidental.

For the sake of reliability of an assessment, which primary goal is to sort students in terms of pass and fail, it is important that this error of measurement is as small as possible, so that based on actual capability, a student passing should actually pass and a student failing should actually fail.

The Standard Error of Measurement (SEM) is used to find the confidence interval for individual students with which can be ascertained that the pass or fail they have achieved reflects actual capability. It is calculated like this, in which x is the achieved test score, SD is the standard deviation and α is the reliability coefficient (Cronbach's alfa or KR-20):

$$\mathrm{SEM(x)} = \mathrm{SD(x)}\sqrt{1 - \alpha}$$

From here, you can calculate the 68% and 95% confidence intervals:

Table 10.
Confidence intervals of a test score, based upon the standard error of measurement (SEM)

Certainty	Confidence interval
68% (used most often)	[test_score – 1*SEM, test_score + 1*SEM]
95%	[test_score – 2*SEM, test_score + 2*SEM]

The confidence interval indicates that if we repeat the measurement for an infinite times in the same circumstances, the average grade (and hence the *true grade*) will be within the 68% confidence interval in 68% of the cases, and within the 95% CI in 95% of the cases. That is, if the circumstances stay the same, i.e. the student does not get tired, anxious, bored etc.

For example, if the student scores 26 out of 50 points, with a cut-off score of 28 (i.e. 5.6 rounded to a whole grade: 6) and the SEM is 5, the 68% confidence interval is 21 to 31 in points (and 4 to 6 in rounded grades, 4.2 to 6.1 unrounded). The student will get a 5 (5.2 unrounded). This means that the student has failed, but maybe should have passed based on his actual capacity and wasn't able to because of the either systematic or accidental error of measurement. The 95% confidence interval is even wider: 16 to 36 points, corresponding to a grade between 3 and 7 (3.2 and 7.2 unrounded).

The standard error of measurement can be used to determine for which students the pass-fail decision might be incorrect. These are students with test results that are closer than one (or two) SEM to the cut-off score (the minimum number of points needed to pass the test). If a test is quite unreliable, you would need to gather more information (e.g. more questions, more test results) to base your grades on.

The uncertainty of grades is a reason to allow for compensation between partial grades within a course, and in the end maybe also between courses.

For example, programmes could decide that students who received a 5 for Dynamical Systems 1, but got a 7 for Dynamical Systems 2, could still receive a 'pass' for Dynamical Systems 1 (or at least not have to take a resit for Dynamical Systems 1 in order to graduate). Especially if the learning objectives of the second course use the ones in the first course.

1.18. Frequency distribution of grades

You can represent a **frequency distribution** of the grades in either a **table <u>or</u> graph**, like in the example below. If you use Contest, use the cumulative frequency distribution graph instead.

You will use the frequency distribution/histogram to discuss whether or not to increase the grades. This might depend on the percentage of students that passed the course. In order to pass the course, students need get a grade of at least the minimum pass grade (in general, 6.0 is used as minimum pass grade, see 0).

In case we study the score instead of grades, students need at least the cut-off score in order to pass the course .

The example is an exam with 70 participating students who answered 15 (sub)questions. The *minimum pass grade* in this example is 6, and the *cut-off score* is a score of 7 points.

Table 11.
Example frequency distribution of grades

Grade	Score	Number of students	Percentage	Cumulative percentage
1.0	0	0	0%	0%
1.5	1	0	0%	0%
2.0	2	0	0%	0%
2.5	3	0	0%	0%
3	4	1	1%	1%
3.5	5	7	10%	11%
4	6	3	4%	15%
4.5	7	20	29%	44%
5	8	15	21%	65%
5.5	9	9	13%	78%
6	10	5	7%	85%
6.5	11	6	9%	94%
7	12	2	3%	97%
7.5	13	2	0%	100%
8	14	0	0%	100%
8.5	15	0	0%	100%
9	16	0	0%	100%
9.5	17	0	0%	100%
10	18	0	0%	100%

This example table tells you that 78% of the students failed. You could use the cumulative percentage in this table to determine a relative cut-off score (i.e. if you think that around 70% of the students should have passed, you could shift the cut-off score from duet.

A relative cut-off score is the minimum score at which a predetermined, desired percentage of students passes the course (or test).

In this case there is a strong indication that your test may have been too difficult and there might be a problem with validity. If, after critically going through the entire analysis, this is proven to be the case, you can use this table as a tool to assess your pre-determined cut-off score. You could for example state that 50% of the students should pass the test. In that case, you could use 8 points as the cut-off score (44% of the students would fail the test).

Putting the **frequency distribution** into a **histogram** will show you if the distribution is normal or whether there is a ceiling or a floor effect. When you have a floor effect, most students have a relatively low score, meaning the test was too difficult for this group of students. When you have a ceiling effect, most students have a relatively high score, meaning that the test was too easy for this group of students.

Examples of both are shown below:

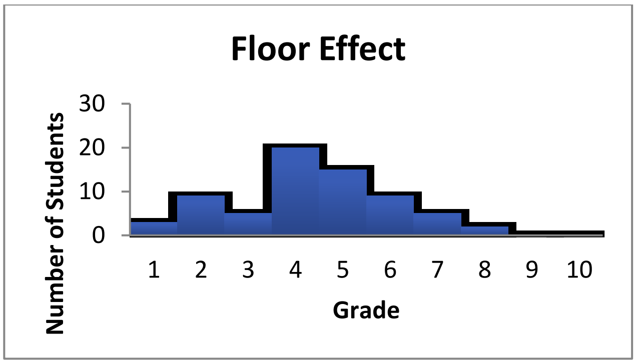


Figure 5. Grade histogram demonstrating the floor effect

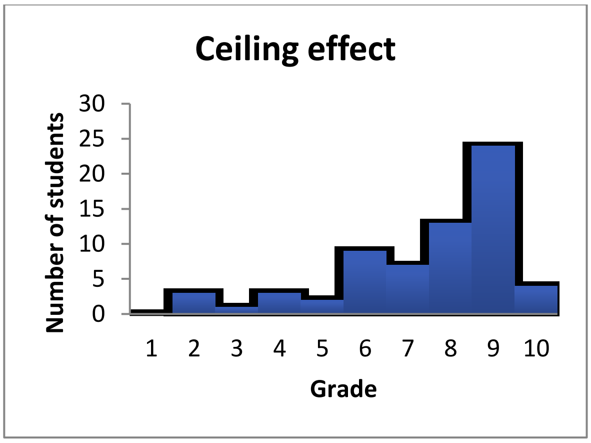


Figure 6. Grade histogram demonstrating the ceiling effect

1.19. Analysis of the quality of the test items

In this section, you will learn the analyse the quality of the individual items. You can use this to pick the most worrying items that you can have a closer look at. You can use this information to change the scoring of the item for the students who just took the exam, and to help you select how you are going to further improve next year’s test.

Keep in mind that it is practically impossible to make flawless assessments (unless we had unlimited time).

Therefore, we must be prepared to make adjustments in the answer model or rubric grading after the test result analysis.

Furthermore, keep in mind that assessments test the extent to which students master the learning objectives. How do students master learning objectives? By engaging in learning activities. We assume that by engaging in learning activities, students will get better at all learning objectives, and that we measure the final attainment during an assessment. This implies that the 'flaws' that you encounter are possibly not caused by the quality in the questions, but possibly partly by the quality of the learning activities.

You can make use of a combination of variables to choose which (for example four) items are worrisome. These three variables are the following:

- Maximum score
- Average score
- Correlation with the other scores

In the following sections, these values are first discussed individually. After you comprehend what kind of information the individual variables can reveal, we will discuss how you can use their combination to focus your attention on potential problems (and solutions).

1.19.A. Maximum score achievement

The goal of our course was to facilitate our students to master the learning objectives, and the goal of the assessment is to measure whether we and they succeeded. For each individual item, we expect that there are students who get full score (if we have a reasonable number of students).

If this is not the case, there may be problems with the answer model, or with the course (learning activities):

- For exams: Will students who master the applicable learning objectives be able to give the model answer, after reading the question? Or could the question lead to other, valid answers that are currently not rewarded?
- For assignments/projects: is it feasible for good students who took your course (taking into account both the available time as well as the learning activities, supervision, feedback, material, assignment instructions and rubric/assessment sheet) to obtain the maximum level for the criterion?

We will call the maximum score *maxa,* expressed in points.

1.19.B. Item difficulty / average score (p)

p is the average, normalized score and has a value between 0 (no points) to 1 (full score). The higher *p*, the higher your students scored on this item, and the *easier* the question or the criterion. For closed questions, *p* equals the fraction of students who answered the question correctly. To summarize: *p* is a reverse measure for the difficulty of an item.

$$\mathrm{p} = \ \text{Average score⍰Maximum score⍰}$$

The complete formula for calculating the p-value is:

$$p_j = \frac{\sum_{i=1}^{N_{stud}} s_i}{N_{stud} \cdot S_j}$$

With p_j the p-value for subquestion j, N_{stud} the total number of students, S_j the maximum score of subquestion j, and with s_i the score of student i on subquestion *j*.

Ideal value: When designing an exam, you would want to include questions that cover a wider range of difficulty, so that the test can distinguish between good and very good students, as well as between pass and fail students. Note, poor performing students refer to those students who did poorly on the assessment overall, while good performing students are those who received a good grade for the entire assessment.

For open-ended questions, the optimal *p*-value is in the range between 0.4 and 0.6 (See 1.11.E 'How question difficulty ' for considerations to deviate from the ideal value of *p*). Although the 'ideal' value of *p* may be 0.5, you don't want your students to on average get 50% of the points.

Please note that *p* in test-result analysis is not related to *p* as in *probability* in statistics. The *p* in test-result analysis has a *p* that stands for *proportion*, not *probability*.

In case of MCQs, *p* are ideally halfway between the guessing score (1/(number of options)) and 1 (see

Table 12). Some programs like Contest also calculate a *p* that is corrected for guessing (p'), meaning that a p' of 0 is defined as the guessing score.

Table 12.
'Ideal' p-values

Number of options	Guessing score	Ideal *p*-value	Ideal *p*-value with correction for guessing
2	0.50	0.75	0.5
3	0.33	0.67	0.5
4	0.25	0.63	0.5
5	0.20	0.6	0.5

***p* below guessing score:** In case of closed-ended questions (MCQs), p-values below or around the guessing score (1/number of options, see

Table 12), this might indeed have been caused by guessing, for example because the topic was not included in the course. If *p* is lower that the guessing score, there either is a misconception amongst students, or another option might be the correct answer instead.

Note: See 1.11.E 'How question difficulty ' for considerations to deviate from the ideal *p-value*.

Extreme *p*-value (either close to 0 or close 1): This may indicate that the question is either too easy or too difficult.

1.19.C. Item discrimination (R_{iR})

Item discrimination is the ability of an item to distinguish between good and poor performing students. If the item discrimination is high, good performing students answer the question correctly and poor performing students answer the question incorrectly.

There are two item discrimination coefficients: R_{it} and R_{ir}. You can always use R_{iR}, but not always the R_{iT}.

Keep in mind that discrimination may be low if the item could be improved, but also if engaging in the learning activities did not contribute to getting a high score on this item. Either students already knew/mastered this before entering the course, or they did not get enough/effective learning activities during the course.

The capital R stands for 'correlation' (referring to Pearson's correlation coefficient ρ) and 'it' stands for item-test, while 'ir' stands for item-rest. Both measure the correlation between the item score and the total test score, or how closely the item measurement resemble the test measurement.

R_{it} measures the correlation of the item score with the entire test score. R_{ir} measures the correlation of the item score with the score on the entire test, minus the item score itself. This is useful when you have a test with fewer than 25 questions, because you prevent the item from correlating with itself. In case of more items, the difference between R_{it} and R_{ir} will be low. The R_{ir} score can be seen as more reliable (less biased), especially when some subquestions have a larger than average weight for the final grade.

The Rit of subquestion is calculated using the following formula:

$$Rit_j = \frac{\sum_{i=1}^{N_{stud}} (x_i - \mu)(s_i - \mu_j)}{\sqrt{\sum_{i=1}^{N_{stud}} (x_i - \mu)^2 \sum_{i=1}^{N_{stud}} (s_i - \mu_j)^2}}$$

With N_{stud} the total number of students, x_i the final score of student *i*, and μ the mean final score, and with s_i the subquestion score of student *i* on subquestion *j*, and μ_j the mean score on subquestion *j*.

The Rir of subquestion *j* is calculated using the following formula:

$$Rir_j = \frac{\sum_{i=1}^{N_{stud}} \left((x_i - s_i) - \widetilde{\mu_j}\right)(s_i - \mu_j)}{\sqrt{\sum_{i=1}^{N_{stud}} \left((x_i - s_i) - \widetilde{\mu_j}\right)^2 \sum_{i=1}^{N_{stud}} (s_i - \mu_j)^2}}$$

Where $\widetilde{\mu_j}$ is the mean test score calculated from all subquestion scores minus the score from subquestion *j*. The R_{ir} and R_{it}-values are always between -1.00 and +1.00[3]. These values can be interpreted as follows in case of closed-ended questions:

Ideal values: We aim at items with a R(it)/R(ir) of at least 0.20 (see

[3] R_{ir} squared equals the percentage of variance in the final grade that is explained by the score for the item. So if R_{ir} of question 4b equals 0.5, it indicates that 25% of the variance of the final score (i.e. the grade) can be explained by the score of question 4b, if we assume a linear relation.

Table 13). Note that these values are less reliable when less than 50 students took the test.

For open-ended questions, projects and assignments, the correlations tend to be much higher. It is wise to always look at the lowest R_{iR}-values of a test.

Table 13.
Interpretations of Rir and Rit values

R_{ir} and R_{it}	Item discrimination quality
0.40 and higher	very good
0.30 - 0.39	Good
0.20 - 0.29	mediocre, the question should be improved
0.19 and lower	bad, the question should not be used or altered completely
Negative values	bad, good students have answered the question incorrectly and vice versa.

Negative values: In case Rir is quite negative, this indicates that overall well-performing students performed worse on this item. It might have been a trick-question, which they have overthought. Or if p is low, only bad performing students seem to have given the correct answer. A multiple-choice question with a low Rir might be an indication that the answer key (answer model) is incorrect, or that there are multiple correct answers.

Value near zero: In case the Rir is near zero (below 0.2), the score for this item is not correlated with the overall score of the other items. In other words, the score on this item does not give information on how well they do in the course.

1.19.D. Attractiveness distractors MCQs (a)

For MCQs only: determine the **quality of the distractors** (the incorrect answer options) by calculating the *a*-value**.** This will give you the proportion of students who choose a particular distractor, and must be calculated for each distractor.

The formula for calculating the *a*-value is:

$$a_k = \frac{N_{stud,k}}{N_{stud}}$$

With a_k the a-value for distractor k, $N_{stud,k}$ the number of students that chose distractor k, and N_{stud} the total number of students.

For each item, the sum of *p*s (proportion of students who picked the right answer) and the *a*-values (proportion of students who picked each of the distractors) is equal to 1.

Ideal value: Ideally, the *a*-values should be about the same for each distractor, because distractors should be equally plausible.

Plausibility distractors: If one of the *a*-values is much lower than the others, that option is not plausible for students, which increases the guessing score. The option could be rewritten, or removed. Formulating plausible distractors is time consuming and very difficult and should not be underestimated. Setting MCQ tests are, therefore, not an 'easy way out'.

Problems with key: If an a-value is higher than a p, students might have chosen the distractor because it was the *key* (correct answer) after all, or because it was a trick question. A relatively low a-values (compared to the other a-value) indicate that an distractor was not attractive enough. Of course, when 90% of students correctly answer a question, the a-values can never be high and in case of low number of students, you cannot draw strong conclusions.

1.19.E. Finding the most worrying items

As discussed previously, the most important indicators that you might need to change the answer model are that

- (almost) no students got the maximum score,
- negative or relatively low Rirs.
- low ps, and
- high a-values (for closed-ended questions).

In order to select the most worrying items, you analyse the combination of these indicators, in the order of importance that is indicated below.

Indicator for worrying item	Implication
1) Maxa < max	None of your students got the maximum score. Was it possible for them to achieve the maximum score, judging from the question, the model answer, and the learning activities? You might conclude that you want to adjust the answer model.
2) Rir < 0 (e.g. -0.2)	Good students performed not good on this question, and/or not-so-good performing students performed good on this question. This is always problematic. In case p is small, this indicates that the few students who answered the question correct, were the bad-performing students. In case p is large, this indicates that the students who answered the question incorrectly, were the good-performing students. Maybe the question was a trick-question, that was overthought by the good students?
3) Rir ~ 0.0 (<0.2) or for open questions: the lowest Rir-values	This question was not good at discriminating between good performing and bad performing students. Assuming that performance depends on course participation, the item did not give information on whether or not students actively participated in the course, which is not ideal.
4) a-value < p-value (MCQ)	This alternative was chosen more frequently that the correct answer. Especially if the Rir is negative, this might be an indication that the key is incorrect.
5) p-value small	Only few students got this question correct. If the Rir is high (relatively), it is 'just' a difficult question, that was only answered correctly by good-performing students, which can be fine. Unless the whole test has low p's and many students failed.

Whenever you have few students, you cannot draw strong conclusions. In general, whatever the grades tell you, you know what happened in class and might have ideas on what is going on.

1.20. How to adjust grading using the test result analysis

In the last section, we discussed how you can identify the most worrying items using a combination of indicators of the test result analysis, and gave you some hints on what the underlying problems might be. In this section, we will discuss how you can adjust the scoring of the items for the students who did the test. Furthermore, if the grades or passing rates are low and not representing the level of LO mastering after adjusting the scoring, we will discuss how you could change the grading.

○

1.20.A. Find indications to adjust the scoring via the answer model

It is important to keep in mind is that it is impossible to make perfect exams, even after thorough peer reviews. On the other hand, you are the expert of the course, and you may have perfectly good reasons not to take actions; as long as you can justify your decisions.

For example, if your exam consists of calculating questions, and of essay questions, students who are good at calculating, might not have good writing skills, and vice versa. This will decrease the Rirs and Cronbach's alpha, without implying problems with the exam questions at all. However, you might consider offering extra exercises for students who are less skilled in calculating, and exercises for those who are less skilled in writing good essays.

When considering to adjust the grading, you always start by considering to adjust the answer model on item level. Only if this does not have the desired effect and if you consider it justifiable, you adjust the calculation of the grade.

1.20.B. Troubleshooting scoring in exams: Adjust the answer model

The first thing you will do is to consider whether the answer model needs to be changed on item level. This can be justifiable if the question was unclear and does not lead to the current model answer, or when the question was too hard or was not aligned with the learning activities and you consider giving partial answers full points. In order to make this decision, you first need to find the cause of the problem. Ask yourself the following:

- Will the question lead to the model answer for students who master the applicable learning objective(s), or are there other, valid answers?
- Was the question clear to the students? Or was it a trick question or could the student interpret it as a trick question?
- Is the model answer correct?
- In case of closed questions: does the question assess only one learning objective at a time?
- Exams: Was the question part of the learning objectives and of the to-be-studied material?
- Assignments: Is the rubric evaluating students on skills that are not related to the learning objectives (i.e. writing/grammar)?

To a certain extent, any answer that answers the question correctly should be granted full points.

For **example**, if you asked 'Explain whether theory B is applicable to the case?', and the student came up with a plausible answer that you did not think of, you can add it to your answer model. Another example: if the question is 'What is the length of beam A?', and you expected your students to write down the whole, lengthy calculation, but did not ask for it, you should grant full points to the question, even if you are not sure whether this student used the correct calculation.

1.20.C. Troubleshooting scoring in assignments / projects

As for exams, for the 'troublesome criteria' in assignments and project, check whether it was feasible for good students who took your course (taking into account both the available time as well as the learning activities, supervision, feedback, material, assignment instructions and rubric/assessment sheet) to obtain the maximum level for the criterion?

For assignment/project criteria, you might consider the following to get ideas on how to improve or develop a rubric (or other answer sheet):

- High RiRs:
  - Are the criteria overlapping? In that case, you might consider reducing the number of criteria.
  - Are graders assessing the individual criteria separately, or do they use their experience and do the refrain from providing information per criterion?
    - Is the rubric user-friendly enough to motivate the teachers to use it?
    - Is the rubric using the same terminology that you are using when discussing student performance?
  - Furthermore, sometimes relatively high Rirs might indicate that too many items are measuring the same thing. You might con-

> sider calculating the correlation between all individual items to check whether this is true.

- Low Rir?

> - Is this criterion measuring something different from the other criteria?
> - Do students who follow the course also practice on this criterion and get LO-oriented feedback on their performance?

- No maximum scores?

> - Is the maximum level realistic?

- Small spread/standard deviation?

> - Is the formulation of the descriptors in the level such, that you can give students high and low points per criterion, or are fail-levels describing levels lower than entrance level?
> - See also 'no maximum score'.

1.20.D. Excluding items or giving students full credits

When considering to exclude questions or criteria from grade calculation by for example giving full points to all students, you have to make a trade-off between the following factors:

- **Validity**: deleting a question or criterion (for assignments) will diminish the representability of your exam of the learning objectives. Reflect on if you have enough questions left per learning objective (and level) for the validity of your exam or assignment.
- **Reliability**: deleting a question or criterion that has a low or negative Rir-value will improve the reliability of the grade. That is, the grade is probably a better reflection of the level to which the students master the learning objectives that were measured in the course.
- **Fairness**: consider whether simply deleting it is fair for all students. Is it probable that students spent a lot of time on this question or criterion? Consider giving students who correctly (guessed?) the answer a bonus point, or giving everybody full grades, although both options will diminish the reliability of the grade.
- **Transparency**: in order to provide transparency, you will need to communicate the change in test grade calculation to the students. If you feel reluctant to do so, it might be because of fairness issues. Because of fairness and transparency, it is not advisable to change the weighing/division of points between questions /criteria afterwards: students who might have put a lot of time in a criterion/question with a high weight, will be disadvantaged if the weight diminishes.
- **Constructive alignment**: Is this question/criterion part of a learning objective? Are you sure that your students had enough possibility to *practice* with this type of question/criterion? Did the students get *feedback* on their performance level on this question/criterion during the course? If one of these question results in a 'no', you could remove the question.

1.20.E. What if the grades are too low?

It should be possible for at least some of your students to score a 10. So, what to do if all the grades are too low? If there was a mistake on the test, or if a question was too vague, you probably already adjusted the answer model. If you still think that the grades do not represent how well students master the learning objectives, you might want to adjust the grade calculation.

> It might be a good idea to check the assessment policy whether you should discuss changes in grade calculations that are based on the test result analysis with your Board of Examiners (since they have given you the mandate to grade students), your programme director and/or the educational advisor of your faculty.
>
> There are several ways to adjust the grading. The most simple one is to simply add a constant number to the grade. Another way is the *Cohen-Schotanus adjustment*. This one is described below.

1.20.F. Cohen-Schotanus' adjustment of score-grade transformation

Cohen-Schotanus (University of Groningen, Medical Faculty) explains that because we can (and often do) make mistakes with our exams (and courses), it is possible to underestimate students' abilities. In short, she assumed that the top 5% of the students are supposed to get a 10. Therefore, she author calculates the *average score of the top 5% students* and assigns them a 10. In line with this method, determine the knowledge percentage and use that to find the cut-off score (after correcting for the guessing score).

The following **example** is the procedure is for a multiple choice exam with 60 questions of 1 point each.

- Total number of points = 60
- Average score of the 5% best students = 55 (example)
- Correction for guessing = 60/4 = 15
- Average *corrected* score top 5% - correction for guessing = 55 – 15 = 40 → students get a 10
- Knowledge percentage = 60% (example)
- Cut-off score = 15 + 0.6*(55-15) = 39 points. Students that have 39 points and more will get a pass.

The Cohen-Schotanus method is only meant to correct grades in large, 'normal' student populations. For retakes, you have a sample of students that is likely to score lower than the whole student population. Therefore, you cannot do a Cohen-Schotanus correction.

1.20.G. Regulations for changing grade calculations

It is good to check in your regulations for whether your faculty has specific advice on how to determine the cut-off score before and after delivering an exam to your students. For example, 3mE uses an Angoff method to determine the cut-off score *before* delivering the exam by estimating how many points the students, who are performing at the minimum pass-level (the level of a 6), will get for each item. After analysing the exam results, the cut-off score is adjusted using the Hofstee method. After this, the examiner can decide to apply a version of the Cohen-Schotanus method to make sure that the student(s) with the highest score will get a 10.

1.20.H. What if Cronbach's alpha remains low after adjustment?

If Cronbach's alpha stays low after having adjusted the answer model, the assessment most likely does not have enough (sub)questions for a valid analysis, and so you do not have enough information to estimate reliably the students' grades.

Another explanation of a low reliability may be that your course assesses different skills, for example, writing skills and calculation skills. As mentioned previously, students who have good writing skills might not be performing well when doing calculations. Could you customize the learning activities to improve 'writing skills' for some students, and 'calculation skills' for other students?

Designing good assessments has four stages:

- Making a blue print (a schematic overview)
- Writing the test itself
- Writing an answer model/rubric
- Getting feedback on step 1, 2 and 3 from peers

For exams (see chapter 0) and assignments, the process is very much alike:

Table 14.
Comparing design process for assignments and exams

	Assignment	Exam
1. Blue print of test	Consistency check table Rows: LOs Columns: deliverables Cells: criteria and weighting	Assessment matrix Rows: LOs Columns: levels of Bloom Cells: (sub)question number(s) and weighting
2. Test	Assignment description	Exam (including front page)
3. Answer model	Answer model - Rubric (or assessment sheet) - Instruction for graders	Answer model - Model answers - Points to be awarded in each situation - Instruction for graders
4. Peer feedback	Peer feedback	Peer feedback

One characteristic of assignments is that the assignment simulates a situation in the work field, and that learning activities and assessment activities are combined into one. Therefore, one could consider that assignments should be constructively aligned within themselves (see Figure 7).

In case of an assignment, this triangle consists of objectives, tasks/instructions and the assessment criteria. These should be aligned. Furthermore, you must make sure that students get feedback on each and every criterion in some form, before they deliver their final product. The feedback might consist of feedback on an early version of a report, or on a pitch, but also on separate exercises (that focus on one or more criteria), or even peer feedback. As long as the students have a reliable indication of on what level they are performing per criterion.

-

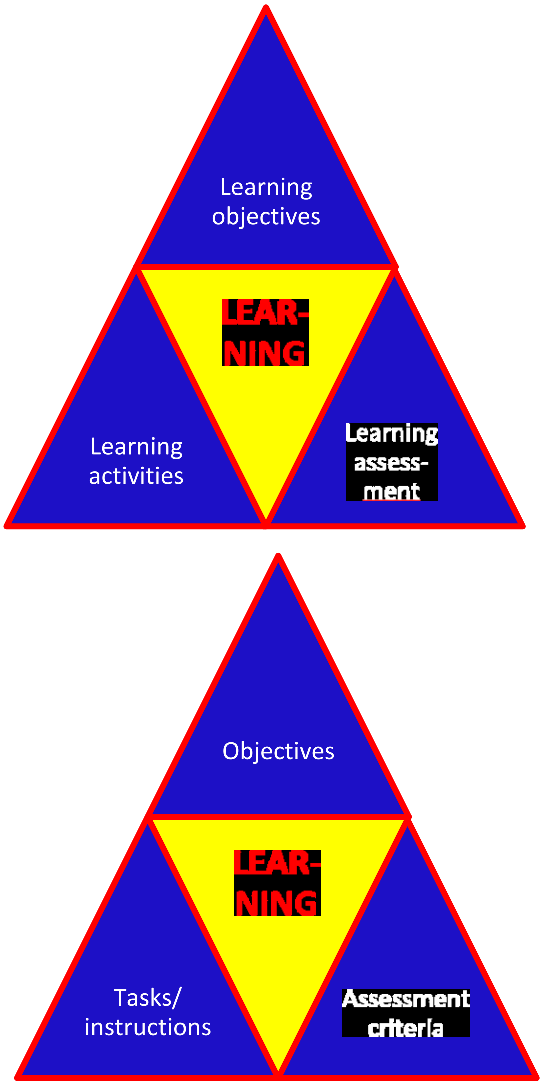


Figure 7. Constructive alignment triangle for a course (top) and for an assignment (bottom)

1.21. Assignment blueprint: consistency check table

In chapter 2 in *Designing assignments used for assessment* (Van de Veen. 2016), an explanation and a step-by-step tutorial of how to design the blueprint of an assignment is explained. This results in an assignment specification form (Figure 2.5, Van de Veen, 2016, pp. 38-39, and a consistency check table (p. 56). The goal of this table is to enable us to check whether:

- All learning outcomes are fully covered by the criteria;
- The division of points between the criteria matches the importance of the criteria and the corresponding learning outcomes;
- Criteria that do not match any learning objective are removed or moved to the 'prerequisite' row, where the knock-out criteria are grouped; and
- The amount of supervision is appropriate for the learning objectives.

Following is a slightly simplified version of the assignment specification form and consistency check table. This is an example of a consistency check table for an imaginary project where students have to design a foot-bridge over the Schie canal in Delft that can withstand a hurricane for first year mechanical engineering students. Each column represents a product that they need to deliver, or in each cell, you can find the **criteria** that they will be assessed on.

Table 15.
Example consistency check table for an imaginary 1st year bachelor project in which students have to design a foot-bridge.

DELIVERABLE, ATTITUDE, SKILL, BEHAVIOR **LO**	**Pitch (group, 0%)**	**Presentation (group, 25%)**	**Report (group & individual, 65%)**	**Contribution (individual, 10%)**	**Total % per LO**
LO1: design a foot-bridge over a canal that meets the operational requirements	Exploration (0%) Considerations (0%) Drawings (0%) Decisions (0%)	Exploration (2%) Considerations & decisions (2%) Drawings (1%)	Exploration (15%) Considerations & decisions (20%) Drawings (10%) Calculations (15%)		65%
LO2: present to an audience of professionals	Presentation technique (0%) Conveying a message (0%)	Presentation technique (10%) Conveying a message (10%)			20%
LO3: work in a group			Reflection on group process (individual, 5%)	Contribution to group process (5%) Contribution to product (5%)	15%
Prerequisites for obtaining a grade			Grammatical and spelling errors do not severely hinder **readability** Use of required **report structure**		

While using a consistency check table, please notice that the columns are called 'tasks' in the book. In general, the columns usually contain the following:

- Deliverables: objects that need to be handed in, for example, a report or a piece of coding; or that has a date at which they are presented, for example, a presentation, poster presentation, pitch);
- Attitudes, skills or behaviours: attitudes, skills and behaviours that are (only) tested during the period that students are working on the assignment or project (e.g. participation, critical attitude, independence, preparation, laboratory skills, programming skills, group work skills).

1.22. Assignment description

Chapter 3 of Van de Veen (2016) discusses how to write a clear and motivating assignment description. Section 3.3 contains very valuable tips. She proposes a format for writing an assignment description, that forces you to include all important parts of such a description. On page 62-63 of Van de Veen (2016), you will find a good example. In 1.26.A 'Checklist for assignments', you will find the checklist that may help you to formulate your assignment.

1.23. Answer model for assignments: rubrics

In an answer model for assignments, you can either use a simple scoring guide rubric (see p. 74 and on this page) or a rubric. More information on rubrics can be found in chapter 4: 4.3 and 4.4, where it is explained what a rubric is and how a rubric can be beneficial. 4.5 to 4.11 will show you step by step how to make a good rubric and which choices you need to make. You will find different types of rubrics in Van de Veen (2016) and can adjust them to your needs. Here is an overview:

- **Scoring guide rubric** (like the one used for evaluating the proof of competence of UTQ module ASSESS) for an essay on the history of the idea of Europe: **p. 74**. This type of rubric only describes the pass level per criterion. Keep in mind that the description should contain the pass level (minimum acceptable level). The example in the book seems to describe a higher level than the pass level.
- **Standard rubric** for a project with a drone that should fly through an obstacle course (deliverables are flight performance of the drone, the program (software), and a report): **p. 77**
- **Three level rubric** of a presentation, including a score and comments column: **pp. 92-93 = pp. 106-107**.

Some additional tips/considerations:

You might want to reverse the order of the columns from 'best' to 'worst' level, so that students can directly read the expectations for the highest level next to the criteria and therefore can quickly determine what is expected from them.

In case your columns are ordered from 'worst', to 'best, the good thing is that you can diminish the text in the descriptors, by making, for example, the 'good' and 'excellent' level build upon the 'sufficient' level. An example of these 'incremental' descriptors, using '…' for the part that is repeated is the following:

- Sufficient: 'Mathematical formulation is correct and variables are individually explained'
- Good: '…in relation to each other'
- Excellent: '…and to the model.'

This helps to keep the rubric simple and clear in a glance. Consider using the rubric for peer feedback for, for example, a draft product. You may replace the grade calculation table by a simple formula, if that suits you better, whether or not you add some minimum levels for all or for certain criteria or criteria groups.

You might (or might not) find it useful to give a better overview by clustering criteria into criteria groups. For example: split the criteria group 'writing style' into the criteria 'clarity', 'conciseness', and 'objectivity'.

One extra tip/consideration about **knock-out criteria**:

- Instead of giving a maximum number of pages excluding figures, you might want to give a maximum number of words, including captions (which makes it easier to check). This might prevent students from using terribly small fonts or placing all figures at the end of their report (making it more difficult to read & grade) to enable them to count the number of pages without figures.

Below, an example of a rubric is depicted for the group-work part of the report of the bridge designing project from the consistency check table (see Table 16).

Table 16.
Rubric for grading the group part of the report of the bridge designing project from the consistency check table in Table 15

LEVEL CRITERIA (%)	Excellent (10)	Sufficient (6)	Insufficient (2)	Score
Exploration (25%)	At least 5 innovative and plausible optioned are clearly described.	At least 4 different options are described, of which 1 is innovative.	None of the described options is innovative; ----- **or** ---- there is a large **overlap** between the options **and** less than 4 individual options can be distinguished.	
Considerations &	The decision is based on a trade-off between quali-	The decision is based on a trade-off between most	The decision is not based on the quality criteria	

decisions (33%)	ty criteria and is based on valid arguments.	quality criteria **and** is the argumentation is mostly valid.	----- **or** ---- the argumentation is missing.	
Drawings (17%)	The drawings provide a good overview of the structure as well as **essential** structural details in a **clear** manner.	The drawings provide a rough overview of the structure **and** some structural details.	Important drawings are missing, or provide no overview of the structure **and** no essential details.	
Calculations (25%)	The calculations are complete and correct.	The main calculations are provided, which only contain minor errors. In case of illogical calculation results, these are detected and discussed.	Crucial steps in the calculation are missing ----- **or** ---- illogical calculation results are not detected.	

As you can see, the knock-out criteria are not included in this example

1.24. Instruction for graders for assignments

When you are grading with a number of colleagues, you will most likely have a meeting (sometimes called 'calibration session') in which you will all grade one or a couple of products (reports, code, etc.) and discuss how you make the grading as objective and uniform as possible and what to do in case you are questioning how to grade a particular criterion or student's product.

For more tips on this, see the exam section on 'Instructions for graders' in section 1.31 on page 57.

1.25. Checklists for assignments

In this section you will find three checklist that may help you to improve your consistency check table, your assignment, and your rubric. Use these to make sure you include everything that has to be included, and to identify opportunities improvement. Keep in mind that some points on the checklist may be more, or less, important for your particular assignment. Furthermore, you probably will have to make a trade-off between practicability on the one hand, and validity and reliability on the other hand.

1.26. Checklist for consistency check tables

Checklist 2.
Checklist for consistency check tables

Checklist for consistency check table	✓
The criteria in the rubric are **the same** as in the consistency check table. *(validity, alignment)*	
The **criteria names** are short, descriptive, specific and clear. *(reliability, transparency)*	
The students get (peer) **feedback** on all criteria first **before being evaluated for a grade** on these criteria. *(effectivity)*	
Each learning objective is fully covered by its criteria. *(validity)*	
The criterion **weightings** are representative of the **importance** of the learning objectives[4] *(validity)*	
All criteria that do not match a learning objectives are prerequisites (**knock-out criteria**; *validity)*	
The criteria are **unique** (no overlap between criteria) *(reliability)*	

1.26.A.Checklist for assignments

Checklist 3.
Checklist for assignment description

Checklist for assignment description	✓
The students are addressed directly. ('you will' instead of 'the students will') *(effectivity)*	
The lay-out is clear. (e.g. use of bullets for steps, highlighting what is important) *(effectivity, transparency)*	
Resources are (literature, formats, example code, etc.) provided, if finding/creating them is not part of the learning objectives. *(validity, effectivity, practicability)*	
The **assignment** is written **clearly** and **concisely.** *(reliability)*	
All **terminology** is likely to be known to all students. (e.g. no regional/national 'general knowledge') *(reliability)*	
The **assignment is aligned** with the learning objectives. *(validity)*	
There is enough **time** to complete the assignment *(practicability)*	
The assignment will lead to a **product** that will **demonstrate** the **level of mastering the criteria** *(validity)*	
The assignment description contains each of the following elements *(effectivity)*: - **introduction**: stating the relevance of the assignment. - **learning objectives**: stating what the student will learn. - **instructions**: explaining the activities that need to be undertaken. - **product**: describing what the concrete result are. - **feedback/evaluation**: criteria for assessment, <u>and</u> when and how feedback will be given.	

tion 4b.

[4] Henk van Berkel, *Zicht op toetsen*, 1999, Van Gorcum, pp 152-153.

[4] To get a more reliable evaluation of how well students perform on important criteria, it is actually good practice to split important criteria into (sub)criteria. This will also give your students and you more information on what aspects of the 'big' criterion students will need to work on.

1.26.B. Checklist for rubrics

Checklist 4.
Checklist for rubrics

Checklist for grade	✓
It is clear what the **weightings** of the criteria are.	
It is clear how the **grade** is derived.	
Performance at the minimum **level of a pass** leads to a **pass grade.**	
It is possible to get a **10**, judging by the criteria descriptors.	
Checklist for descriptors	✓
It is feasible to get a **10**, judging by the **descriptors of the highest levels**.	
The descriptors are **objectively** formulated (no 'just sufficient' 'excellent').	
The descriptors are **specific** and **clear**.	
The descriptors of each criterion are **unique** (**no overlap** between descriptors of adjacent levels).	
Checklist for usability	✓
The rubric gives a good **overview at first glance** (not to many rows or columns).	
The rubric fits on one A4.	
The lay-out is clear.	
The **amount of details** is suitable (not too detailed / no information that belongs in a course book).	
There is **space** for specific (individual) **feedback**.	

1.27. Group skills: to assess or not to assess?

If you have decided to have your students do your assignments in groups, there are two questions that we need to answer:

- Do we assess the students on soft skills like 'group skills'?
- Do we train them on group skills?

Even if you decide that you do *not* want to assess group skills, group performance may be limited by problems with group skills. Therefore, group skills will influence the grade, whether you like it or not. This will limit the validity and reliability of your grade. And more importantly, it might hinder learning. Not all students naturally possess group-work skills. Therefore, they need your help, feedback and guidance.

Here are some common subjects that group members might have different opinions on, which will negatively influence group performance:

- Levels of ambition (for example the desirable grade),
- Communication standards,
- Collaboration,
- Time needed to complete the work,
- Working hours
- Choosing a place to work,
- Decision making, and
- Problem solving.

You can have your students discuss these things openly during a kick-off meeting, and to reach an agreement before starting the project. You can have the students monitor each other's behaviour using Scorion. They can also give feedback on each other's work using Feedback Fruits and Presto.

If you choose to grade the group process, you can do so on the level of an individual, or on the level of the group. In both cases, you must make sure that you have enough observations to base your grade on. For individual grading, you might grade the student's behaviour in the group, her evaluation of her group's behaviour, and the quality of the student's own skills, needed for the project. You can also evaluate at the group level yourself, or give the group responsibility for this process. In that case, you could evaluate, for example:

- The product/content (product, report, presentation, interview, portfolio, customer evaluation),
- The process/planning (project plan, planning, logbook, criteria list, study contract, portfolio, report), and
- The cooperation (evaluation report, individual reflection report, criteria list, process report, presence list, peer evaluation).

Designing an assessment has four stages:

- making a blue print of the test (a schematic overview)
- writing the test itself
- writing an answer model
- getting feedback on step 1, 2 and 3 from peers

For exams and assignments, the process is very much alike:

Table 17.
Comparison of assignments and exams

	Assignment	Exam
1. Blue print of test	**Consistency check table** Rows: LOs Columns: deliverables Cells: criteria and weight	Assessment matrix Rows: LOs Columns: levels of Bloom Cells: (sub)question number(s) and weight
2. Test	Assignment description	Exam (including front page)
3. Answer model	Answer model - rubric (or assessment sheet) - instruction for graders	Answer model - model answers - points to be awarded in each situation - instruction for graders
4. Peer feedback	Peer feedback	Peer feedback

1.28. Exam blue print: assessment matrix

1.28.A. What is an assessment matrix?

An assessment matrix is a blueprint to help you check whether your assessment covers the learning objectives you set and whether you test at the right level of thinking skills (the validity of your course). It is sometimes called a 'specification table' as well.

You can make an assessment matrix on course level and on test level. This document discusses how to make an assessment matrix for a single test. Assessment matrices can be used for exams that consist of individually graded questions, like written exams, oral exams, or practicals in which students have to answer a fixed set of questions (as opposed to writing a report). This document explains in detail how to make an assessment matrix.

The module you teach has a set of learning objectives or learning outcomes. In this course, we assume you use Bloom's taxonomy to define those learning objectives. The learning objectives for a course could look like this for example:

On successful completion of the course, you will be able to:

1. List and define basic reliability, availability, maintainability and supportability (RAMS) concepts and measures.
2. Describe the main elements necessary to perform maintenance modelling and analysis for aerospace applications.
3. Identify common assumptions in maintenance modelling and analysis.
4. Select appropriate modelling and/or analysis techniques for given problems in the aerospace domain through analysis of problem characteristics.
5. Apply modelling and/or analysis techniques for given problems in the aerospace maintenance domain by:
    a. Formulating and solving stochastic time-to-failure models to determine aircraft system and component reliability characteristics.
    b. Formulating and solving time series techniques and stochastic demand arrival models to determine and predict aircraft system and component supportability characteristics.
6. Evaluate the benefits and drawbacks of available options for modelling and analysis of a given problem in the aerospace maintenance domain.

Figure 8: Example set of learning objectives for a module

An objective specifies a topic or a bit of content (such as *RAMS concepts*, or *stochastic time-to-failure models*) as well as what the student should be able to do with that topic (*list, describe, apply).* The verb indicates the intended level of Bloom's taxonomy that this objective aims at. In this example the first objective (*list/define)* is aimed at the bottom level (remember), whereas the final objective is aimed at the highest level of evaluate.

To develop an assessment (an exam or an assignment) that is representative of these objectives, these two aspects, topic and level, both need to be taken into account. This is where the assessment matrix comes in. Basically, it is a table in which the two aspects of the objectives are related to the parts of the test, yielding a convenient overview of the composition of the test.

The matrix shows how the test is composed. What is the contribution of each objective towards the final mark? And to what extent are the different levels of Bloom's taxonomy tested? This is convenient for the person creating the test (does it match my intentions?) and also a quick way of communicating the composition of your test to someone else.

An example of an existing exam whose assessment matrix was reverse engineered is given below. In the table, Q is the (sub)question number, and P is the points per (sub)question.

Table 18.
Assessment matrix for an existing exam based on the learning objectives listed previously. Q = (sub)question number, P = points per (sub)question

Learning objective	Bloom's cognitive levels								Total points (% of total score)
	Remember		Understand		Apply		Analyse		
	Q	P	Q	P	Q	P	Q	P	
1	1a 1b 3a	3 4 3							10
2	1c 4	5 5							10
3			1e 2a	3 5					8
4			1e 2b	5 5	2c	5	3	5	15
5a			1d	7	1e 4	5 5			17
5b			3	5	3	10			20
6							3 4	10 10	20
Total	20		30		25		25		100

1.28.B. Constructing an assessment matrix for a new exam

By following the steps below, you will first design an assessment matrix which shows how you would like to construct the next exam. Then, you will analyse an existing exam and investigate to what extent it matches your "ideal" matrix.

i. Step 1: List the learning outcomes

Start by listing the learning outcomes in the left-hand column of the test matrix. If there is only one summative assessment, final exam, then all of the learning outcomes of the module need to be included. If the module is assessed in multiple ways (for example, a group-work project and an exam), then you need to select those learning outcomes that you want to test in the exam.

ii. Step 2: Determine the weight of each learning outcome

Now that you have listed the learning outcomes that will be tested, the next step is to decide what weight you would like each learning outcome to have. In other words, what percentage of the total score should each learning outcome represent? Are they all equally important? Or do you want some outcomes to have more weight in the exam?

Complete the final column of the matrix, by filling in the weighting of each learning outcome.

iii. Step 3: Determine how each learning outcome will be tested

Now that you have decided the weighting of the learning outcomes, you can complete each row of the matrix by deciding at which cognitive levels you want to test each outcome. If formulated correctly, a learning outcome indicates what level of cognitive skill is intended.

For example, suppose Outcome B in the matrix above is the learning outcome “Apply modelling and/or analysis techniques for given problems in the aerospace maintenance domain”. This outcome is at the level of application, and you have decided that it should count for 30% of the total score.

What are your options for completing this row of the assessment matrix? You definitely need to allocate a proportion of the weight to the “application questions” cell, or you would not be testing this learning outcome properly. You cannot test at levels above the application level; that would not be fair.

You could decide to only test this outcome at the application level and put 30% there. However, there are also good reasons for testing a learning outcome explicitly at the level or levels below it. One of them is that this gives you and the student feedback on what level of skill they have reached. Some students might answer the application level questions incorrectly, but have no difficulty with the comprehension questions that relate to the same learning outcome.

Another reason may be that you want to build up the question in steps: first recall the facts required, then apply them to a new case.

So, in this example you might decide to allocate 10% to comprehension questions, and 20% to application questions. Or 15%-15%. Or 5% reproduction, 5% comprehension and 15% application. Or some other combination – it is up to you.

iv. Step 4: Check and adjust the totals for each level

After step 3, add up the percentages in each column to complete the totals in the bottom row. When you have done this, check whether you are happy with the result. You may find that you want to make some adjustments.

For example, if in step 3 you allocated a percentage to the reproduction level for every learning outcome, you may now realise that the total for this column turns out higher than you would want.

If you are happy with the totals in each column, then you are done with designing your assessment matrix. If not, then you need to make adjustments to the cells, until you are happy.

If you are designing a new exam, for a new or redesigned module, then the next step is to start constructing questions that match the matrix. If you have designed a matrix for an existing exam, it is interesting to check how well this exam matches the matrix that you have just constructed.

The assessment matrix will now look something like this (see

Table 19):

Table 19.
Assessment matrix for a new exam

Learning objective	Bloom's cognitive levels						Percentage of total score
	Remember (recall basic information)	**Understand** (explain ideas and concepts)	**Apply** (apply information in a new way)	**Analyse** (distinguish components)	**Evaluate** (justify a stand or position)	**Create** (create a new product)	
LO 1	5%	5%					10%
LO 2	5%	5%	20%				30%
LO 3		20%					20%
LO 4		5%	10%				15%
LO 5			25%				25%
Total	10%	35%	55%				100%

In this example, the number of questions in each cell has not yet been specified. This can be done while you are making the exam, or you can do it now.

You can delete columns you are not using for clarity.

1.28.C. Analysing an existing exam

To what extent does the existing exam match the blueprint that you have just constructed? To figure this out, go through the questions in the exam and for each (sub)question, decide in which cell of the matrix it belongs.

This means that you need to decide which learning outcome it relates to and what the level of the question is in terms of Bloom's taxonomy.

Write down the question number and the number of points that can earned with this question in the appropriate cell. You can add this information to the matrix you have constructed, or you can complete a new one. Here is a template for an assessment matrix:

Table 20.
Assessment matrix for an existing or a newly designed exam. Q = (sub)question number, P = points per (sub)question

Learning objective	Bloom's cognitive levels												Total points (% of total)
	Re-member		Understand		Apply		Analyse		Evaluate		Create		
	Q	P	Q	P	Q	P	Q	P	Q	P	Q	P	
1													
2													
3													
4													
5a													
5b													
6													
Total													100

When you have done this, you can add up the points, convert them into percentages and check to what extent the exam matches the new matrix. If there are differences, what are they? What are the main areas you would want to change (if any)?

Additionally, by adding an extra column to the table that includes the time the students spend in total on a particular learning objective, you can compare the percentage of points to the percentage of hours. By 'hours' we mean 28 hours * the number of EC in your course, i.e. the total time students are supposed to spend on your course. Let us consider an extreme example where students spend 50% of their time practicing LO1, while they only receive 10% of the points on their final exam. If they performed very well during the course on this LO1, this will not have a big influence on their final grade. Furthermore, students might choose not to study LO1, since they will not get much points for this. Therefore, it is wise to align time spent and points given for a certain learning objective.

A few final words about assessment matrices:

An assessment matrix is useful because it provides an overview of the test. Many people find that when they fit an existing test (which they made without using a matrix) into a matrix that the result does not exactly match their intentions, especially with respect to the level of the questions. Often the test turns out to have more lower level questions (especially reproduction level) than intended.

At the same time, it is good to remember that the assessment matrix is an abstraction. It is only meaningful to the extent that the test actually matches the matrix. So making sure that you construct tasks (questions or assignments) that elicit the desired behaviour at the intended level of cognitive skill is paramount.

1.28.D. Number of exam questions

There are some rules of thumb to come up with the number of exam questions.

- The number of questions per learning objective should represent the importance of the learning objective.
- It can be better to have multiple small questions on a learning objective, than one big question. The reason is that you then have multiple 'samples' from a learning objective, instead of a single one. This will improve the reliability. On the other hand, in LO's at higher Bloom levels, it might diminish the difficulty or even the Bloom levels, if you ask a couple of short questions, and one long question might be better for that learning objective.
- The number of points on an exam question must be a good indication of the amount of time students will need to answer the question. Students will try to get the highest grade possible, and will skip questions if they are very difficult and will only result in few points.
- Exam duration: there are some guidelines about how much time it will take a student to answer questions, but this differs quite a lot between type of questions. The best way to determine this is to ask a colleague who teaches a similar course.
- Consider the total number of points in your exam and think about how much the grade will change in case a student misses a subquestion. Will her grade drop from a 10 to an 8? Is that desirable or is the drop to coarse? If not, add more questions in order to make the steps smaller.

1.28.E. Closed questions (e.g. multiple choice questions) and precision

There are rules of thumb for the number of closed questions you need to get a reliable exam. The 'problem' with closed questions is that students can guess a correct answer, without knowing the subject thoroughly enough.

The rules of thumb is:

Single, high stake exam, around 100% of the final grade	
Required Cronbach's alpha	0.8
Number of options	180
MCQ with 4 options	40 questions
MCQ with 3 options	53 questions
MCQ with 2 options / true-false questions	80 questions
Midterm, e.g. 40-50% of the final grade	
Required Cronbach's alpha	0.7
Number of options	120
MCQ with 4 options	30 questions
MCQ with 3 options	40 questions
MCQ with 2 options / true-false questions	60 questions

For a multiple choice exam with 40 questions with 4 answers per question, students will only get higher than a 1.0 in case they have more than 10 questions correct. This is because students will *on average* (some are lucky, some are unlucky) be able to guess 10 questions correctly, without studying for the test. As a result of the guessing correction, the first 10 correctly answered questions will not increase the grade. For the other correctly answered questions, each of them increases the grade by 9/30 = .30.

For an exam with 40 true/false questions (2 answers per question), students will only get higher than a 1.0

in case they answered more than 20 answers correctly. Starting with the 21st correctly answered question, each correctly answered question increases the grade by 9/20 = .45. In case of 80 true/false questions, the precision would be .23.

i. Exam with open and closed questions

In case of an exam which is a combination of open and closed questions that count for less than 50% of the exam, make sure you have at least 80 options, in order to get relevant information from these questions.

1.29. Writing exams and exam questions

The most important hint is to write the exam questions together with the answer model, and use a peer to review them. Have your peer checking whether the question will probably lead to the answer in the answer model, or if the question needs clarification or whether additional instructions are needed.

Below, you will find checklists for the cover page of an exam, for writing exam questions and specific checklists for writing closed and open exam questions, that will help you to formulate and improve your questions and those of your peers.

1.29.A. Checklist for cover page of exam

Some faculties have a standard cover page which is used for all exams. If your faculty does not use one, you can make your own using this checklist. However, not all items might be useful to include in your exam.

Including a cover page may prevent unnecessary stress and loss of points for some students. They can check whether pages/questions are missing from their exam booklet, whether or not it makes sense for them to write essays that hopefully include the correct answer, or if there is anything that they might not be aware of that could diminish their grade.

Checklist 5.
Checklist for exam cover pages

Item	Details to include	✓
General information	- Number of pages - Number of questions - Duration of the exam/start and end time - Course name - Exam date and location - Examiner's name - Name of the second reader/reviewer	
Grade information	- Total number of points - Exam grade calculation and/or cut-off point [minimum points to get a minimum pass grade (6.0)] - In case the minimum grade for this exam is different, for example, 5.0, also mention the number of points needed for this minimum grade - General rating information (if applicable), for example: ○ if (and when) (minor) spelling and grammar mistakes will influence the grade ○ how you will rate a question in case of multiple answers, which are (partly) incorrect ○ how you will rate a question in case redundant information, which is (partly) incorrect	
Instructions	- Resources allowed ○ use of books, readers, notes, slides ○ use of (graphic) calculator, mobile telephones, etc. - Whether name, student number, and programme should be written on all sheets/pages that the student hands in - Whether the number of sheets that the student hands in should be written down (and where) - Any additional information, for example, if certain questions should be answered on separate sheets - Whether students can take the questions, answer sheets or scrap paper with them	

1.29.B. Checklist for validity, reliability and transparency for all types of questions

There will almost always be a trade-off between the quality requirements for assessment, but there are some basics that need to be in place, regardless:

Furthermore, if you have your answer model ready, make sure that the questions will **lead to the answer in the answer model.** This sounds obvious, but it happens all too often because there is a misalignment between what the students should be able to answer/demonstrate, and that which the question requires them to answer/demonstrate. It is easier to pick up on this type of misalignment when you have a complete answer model.

Checklist 6.
Checklist for validity, reliability and transparency

Item	Details to include	✓
Test only one learning objective at a time (validity)	- Do not try to cover more than one learning objective in the same question.	
Relevance of each question (validity)	- Is it clear what knowledge or skill is being tested? - Is this knowledge or skill absolutely necessary in order to answer the question? - Is the answer model in line with what the test questions ask?	
Language (reliability)	- Are there any spelling errors or typos? - Is the question unambiguous and is it clear what is being asked? - Have double negatives been avoided? Is the question concisely formulated?	
Presentation	- Is the layout clear? - Are the figures clear?	
Transparency	- During the test/assignment, are the points to be earned by each question or subquestion announced? This way students can budget their time to be most impactful for them. They should not spend a lot of time on a question that will not earn them a lot of points. - Before taking the test/assignment, do students know ahead of time what will be on the test both in structure and in content? - Before taking the test/assignment, did your students get experience with the types of questions with which you will be testing? - After getting the grade and feedback, does the student get information on how her grade has been calculated, and on how she can improve her performance, for example per learning objective, criterion or subquestion?	

1.29.C. Checklist for closed-ended test questions

Closed test questions can be true/false questions, multiple choice questions, 'fill in the blanks' and pairing questions.

Checklist 7.
Checklist for closed-ended test questions

Item	✓
Make sure the question ends in a question mark. Students should be able to answer the question without looking at the answer[5].	
Check whether the distractors seem as plausible as the correct answer	
Make sure all **options** are roughly of the **same length**.	
Check if a certain question **inadvertently provides the answer** to another question.	
Make sure there are **no grammatical clues** to indicate the right answer.	
Try to **distribute** the right **answers randomly** over A, B, C, D, etc.	
Avoid questions that start with '**Which of the following** statements are true/false?'	

Asking a question like 'Which of the following statements are true/false?' could potentially test more than one thing at a time. If it were an open question, you would have asked and graded the answers to each statement separately with partial points.

All distractors should be equally probable. Constructing the distractors will be a time consuming process. It is better to have more questions with less distractors than having ones that are not probable. As a guideline, use 3 options (i.e. 1 correct answer and 2 distractors). When constructing them, think of the answers that weak students would give if it were an open question.

1.29.D. Checklist for open-ended test questions

Open-ended questions are any questions where the student has to write a free-form answers. The answers can consist of single words, phrases, bullet points, a few sentences or even an entire report.

Checklist 8.
Checklist for open-ended test questions

Item	Details to include	✓
Use a 3-part Structure	- Context (optional) - Question (assignment) - Directions for answering, for example, 'Motivate your answer, showing which formulas you used. Write no more than 3 sentences'.	
Be specific	- Use imperative sentences ("List three characteristics of X" rather than "What are the characteristics of X"). - Specify what you expect in the answer (e.g. "List the <u>three</u> characteristics of X"). - Avoid "anything goes" formulations such as "What do you think…"	
Context and question should be linked	- Make sure the context is relevant for the question. If not, delete it. - If the question can be answered without using the context, then change/remove the context OR change the question. Unless a learning objective is to filter out irrelevant information, of course.	
Check for copy/paste errors	- For example, between old and new questions	

Make sure to have a rubric or answer sheet for grading open-ended questions. This will also help you to keep your assessment aligned with your LOs.

1.30. Exam answer model

Before discussing the answer mode, you must realise that there is a difference between *model answer* and *answer model*. A *model answer* is the ideal answer, that you might want to publish for your students. The *answer model* is a tool that will help you and your fellow graders decide on how to add or subtract points for individual students in a consistent and objective way. It indicates how much points are awarded per correct step or correct part of the answer in case it is based on *addition*, and/or how many points are deducted for all expected if the answer model is based on *deduction (subtraction)*.

[5]

An *answer model* can probably never cover all creative answers that students will come up with. Therefore, you also need an *instruction for graders*, that will tell the graders what to do in these cases. It is advisable to have a meeting in which you discuss difficulties in grading 'creative' or otherwise unexpected solutions, and adjust the answer model accordingly. This might lead to redoing the grading of some of the subquestions.

In section 1.12 on page 22, issues that will diminish the objectivity of grading and hence the reliability of the assessment were described. An answer model enables you to assess the answer as objectively as possible to avoid those issues. The following table gives a checklist of what the answer model should contain:

Checklist 9.
Checklists for answer models

Item	Details to include	✓
The correct, or an ideal answer	- Include all possible answers and guidelines of how much, and how many of these possible answers the students need to give to earn points. - Also give instructions for correct answers that are not included in the answer model.	
The maximum number of points	- Include this both for main and subquestions. - Make sure that the students can earn a reasonable number of points for the amount of work to be produced.	
Description of how divergent answers are marked	- Which answers are considered fully/half/not correct? - How many points do the various half-correct questions still receive?	
Be clear on how interrelated subquestions are marked	- How can points be earned for part-right questions (method)? - Can a student continue calculating with an imagined set of numbers if a first subquestion was answered incorrect? Make sure to instruct your students on this, too!	

Following the checklist when developing answer models can help you avoid potential disputes and increase the overall quality of the assessment.

By developing the answer model at the same time as formulating the question, this can also serve as a check as to whether the phrasing of the question is specific enough. It is a tool that can help make the formulation of the question more pointed, so that the quality of the question is enhanced. If the answer model contains a large number of possible answers, this usually means the formulation of the question is not specific enough.

1.31. Instructions for graders

If there will be several assessors grading the same assessment, an answer model should include general rules for the assessment. Some of these were also mentioned in the previous section:

How to handle subquestions that are mutually dependent (scoring method)?

What to do when the given answer is not included in the answer model or when you are uncertain about the correctness of an answer, for example because the lecture about this topic was given by someone else?

- Will you discuss this with your colleagues?
- How will you add this to the answer model?

The instruction for graders might also describe which other measures you take to increase the reliability, for example to:

- Assess the answers per question (instead of the full examination per student).
- Change the sequence of the students per question.
- Give the students anonymity by having them state only their student number on the answer sheets and not their names.
- Use several assessors per question.
- Divide the different questions over the different assessors, instead of dividing the students over the assessors. In this way, the assessor differences average each other out.
- Grade the first couple of exams together and have a meeting in which you discuss differences between grades and adjust the answer model.

Although it might seem like a lot of extra work, investing time in this can greatly improve the quality of your assessment.

TABLES OF REFERENCE

Table of tables

Table of figures

Table of checklists

References

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher education, 32*(3), 347-364.

Cauley, K., & McMillan, J. (2010). ormative Assessment Techniques to Support Student Motivation and Achievement. *learing House: A Journal of Educational Strategies, Issues and Ideas, 83*(1), 1-6.

Dunn, L. (2002, June 27). *Theories of learning.* Retrieved 11 13, 2018, from Learning and teaching briefing papers series: https://www.brookes.ac.uk/services/ocsld/resources/briefing_papers/learning_theories.pdf

Dunn, L. (2018, November 13). *Selecting methods of assessment*. Retrieved from Oxford Brookes University: https://www.brookes.ac.uk/services/ocsld/resources/methods.html

Garfield, J., & Franklin, C. (2011). Assessment of Learning, for Learning, and as Learning in Statistics Education. In C. Batanero, G. Burrill, & C. (. Reading, *eaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education* (Vol. 14, pp. 133-145). Dordrecht: Springer. doi:https://doi.org/10.1007/978-94-007-1131-0_16

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*, 81-112.

Hulshof, C. (2012, 12 18). *Hoe bereken je cijfers voor een toets?* Retrieved 11 13, 2018, from Blogcollectief onderzoek onderwijs: https://onderzoekonderwijs.net/2012/12/18/hoe-bereken-je-cijfers-voor-een-toets/

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Nightingale, P., Te Wiata, I., Toohey, S., Ryan, G., Hughes, C., & Magin, D. (1997). *Assessing learning in universities.* Sydney: University of New South Wales Press.

Shute, V. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*(1), 153-189.

van Berkel, H. (1999). *Zicht op toetsen. Toetsconstructie in het hoger onderwijs.* Assen: Van Gorcum.

van de Veen, E. (2016). *How to assess students through assignments. A guide to creating assignments and rubrics in higher education.* Voorthuizen: Communicatiereeks.

Wiliam, D. (2011). What is Assessment for Learning? *Studies in Educational Evaluation, 37*, 3-14.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*(4), 477-501.

Index