

TU DELFT ASSESSMENT MANUAL

TEACHING AND LEARNING SERVICES

VERSION JANUARY 2023

Compiled and edited by Lisette Harting, OC Focus, TU Delft, October 2017

Last update: Lisette Harting, Puk Sies, Bouke van Bergen Bravenboer, Janneke Blok (Teaching & Learning Services, TU Delft), January 2023

TABLE OF CONTENTS

<i>Table of contents</i>	2
<i>Introduction</i>	4
CHAPTER 1) Assessment plan	5
1.1. Set-up of the assessment plan.....	5
1.2. Example assessment plan	6
1.3. Constructive alignment of assessment methods	9
1.4. (Dis)Advantages of open and closed-ended questions	11
1.5. Possibility for feedback: formative assessment.....	12
1.6. Digital assessment tools.....	14
1.7. Regulations and guidelines for assessment.....	14
1.8. Quality requirements for assessment	16
CHAPTER 2) Grading	19
2.1. What is a grade?	19
2.2. Grade calculation.....	20
2.3. Objectivity and reliability of grading.....	24
CHAPTER 3) Analysis of test results	26
3.1. Analysis of the achievement of learning objectives	26
3.2. Analysis of the quality of the test items and answers.....	27
3.3. How to adjust answer models based on test result analysis	31
3.4. Reliability of the test (Cronbach's alpha)	33
3.5. Adjusting the grades	35
3.6. How to use correlation table to adjust criteria or rubrics	36
CHAPTER 4) Creating and improving projects/assignments	38
4.1. Assignment blueprint: consistency check table	39
4.2. Assignment description	40
4.3. Assessing assignments: rubrics and grading instructions	40
4.4. Checklists for assignments	41
4.5. Group skills: to assess or not to assess?.....	43
CHAPTER 5) Creating and improving exams	44
5.1. Exam blue print: assessment matrix	44
5.2. Assessing exams: answer model and grading instructions	48
5.3. Checklist for exams.....	50
Tables of reference	54

Table of tables.....	54
Table of figures.....	54
Table of checklists	55
References	55

INTRODUCTION

In this manual, you will find guidelines on how to develop good quality assessments.

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.

The development and finalisation of an assessment consists of several phases. Together, they are known as the test cycle. This manual will guide you through the most important steps of the assessment cycle (see Figure 1).



Figure 1. Assessment cycle

Step 1, the development of an assessment plan, is covered in Chapter 1. An assessment plan includes:

- How you combine formative and summative assessments in your course;
- How your assessments lead to a course grade;
- That your assessment meets the quality requirements for assessment;
- That your assessment meets the rules, regulations and assessment policies that apply to your course.

Step 2, how to design assessments or improve existing ones, is also covered in Chapter 1. This includes designing or improving:

- The blueprint of the assessment (step 2a);
- The assessment instructions (step 2b);
- The assessment criteria (step 2c).

This manual does not cover the process of written exams (step 3). You can find more information on this in the Rules and Procedure for Examinations, which can be found via the link at the bottom of [this page](#) and via your programme coordinator.

Chapter 2 discusses the concept of calculating grades. Steps 4 and 5 of the assessment cycle are covered in Chapter 3. This includes how you can use the test results (4a) to measure your students' mastering of the learning objectives (4b and 5), estimate test quality (4b), and to (re)calculate the test grade (4b and 4c). Chapter 4 focusses on creating and improving projects and assignments, while Chapter 5 provides a similar overview for exams.

CHAPTER 1) ASSESSMENT PLAN

An assessment plan contains detailed information about the *constructive alignment* (the alignment of learning objectives, teaching/learning activities and assessment) 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. You will also see the assessment plan in detail, and read an example of what it could look like.

1.1. Set-up of the assessment plan

To give a good overview of the constructive alignment of your course assessment, include an *assessment overview* (a tabulated summary of the assessment plan). The assessment overview can be included in your Brightspace course for your students to see what assessments they can expect and can be used 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. An example of an assessment plan can be found in Section 1.2.

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

General	
Assessment name	Descriptive name of <u>all</u> assessments (formative and summative)
1. Assessment method alignment	
Assessment method	Examples: midterm exam, homework assignment(s), project, presentation. The method should 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 (0% for formative assessment)
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', 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 Teaching and Examination Regulations (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.

In the table, you can 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), and/or a pass/fail decision.

- The goal of formative assessment is to monitor student learning to provide ongoing feedback that can be used by instructors to improve their teaching and by students to improve their learning. Therefore, formative assessments are assessments that usually do not contribute to the grade of the course. Students should receive feedback on how well they master the learning objectives, and 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.

1.2. Example assessment plan

An example assessment plan can be found in Table 2. The main characteristics and considerations of an assessment plan will be discussed in the remainder of this section, based on this example assessment overview.

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 MCQs 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 MCQs 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.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 LO's, 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 LO's 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. The lecturers have enough TAs and lecturers to do this. Furthermore, students 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 are 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 to a sub-question 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 lecturers will discuss how the questions could have been answered. The lecturers 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.

1.2.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

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.

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. Constructive alignment of assessment methods

TU Delft works with the principle of constructive alignment. 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 need to provide 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. If assessments do not aim towards students meeting the learning objectives for the course, they can be considered redundant.

The following two sections 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 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 a 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 and corresponding assessment methods (Nightingale et al, 1996)

Thinking critically and making judgements	
Developing arguments, reflecting, evaluating, assessing, judging	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - Written examination: - Open questions - Essay questions - Short answer questions - Closed-ended questions: - True/false - Multiple choice - 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	<ul style="list-style-type: none"> - Portfolio - Presentation - Projects - Performance
Accessing and managing information	
Researching, investigating, interpreting, organising information, reviewing and paraphrasing information, collecting data, searching and managing information sources, observing and interpreting	<ul style="list-style-type: none"> - Annotated bibliography - 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	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - 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. (Dis)Advantages of open and closed-ended questions

In general, multiple choice questions (MCQs) in which students have to demonstrate understanding, are very useful in a classroom setting where students can discuss their answers. This can deepen their understanding and analytical skills.

For summative assessment, there are several reasons to decide (not) to use multiple-choice questions.

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 needs 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 MCQs 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 MCQs 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 MCQs than for open questions in order to construct a reliable exam (see 5.1.d, 'Number of exam questions').

1.5. Possibility for feedback: formative assessment

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, you 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) (William, 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 teaching team 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;
- 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 you understand how students are doing in your course and what they will still need (from you) to reach the learning objectives. You can use this information to adjust 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 considered 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. How to go about this will be discussed in detail further on in this manual.

1.6. 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.

1.6.a Central supported digital exam tools

The following **digital exam** tools are centrally supported and follow the archiving regulations for you:

- Grasple (focusses on formative math exams).
- [Ans Delft](#)

For **summative** and **formative digital exams**, Ans is the recommended solution at the moment. Ans is designed to support **paper, digital and hybrid** exams.

For more information and support, contact digitalexams@tudelft.nl for summative exams and Teaching-Support@tudelft.nl for formative assessment.

For a current overview of centrally supported tools, take a look at [this page](#).

1.6.b Other digital exam tools

The following tools are not centrally supported and therefore the course's examiner is responsible for securely archiving and destruction of the data after the retention period has passed.

- Zesje (open source, Latex based)

- [Work2grade](#) (TBM, Pieter Bots)

1.6.c Peer evaluation / feedback tools

There are also centrally supported **peer evaluation/feedback** tools:

- BuddyCheck (to improve behaviour and group dynamics, follow-up of Scorion)
- FeedBackFruits

Reminder: the examiner is responsible for giving the grades!

In addition, the following tools are available for evaluating and giving feedback to student deliverables like reports and presentations:

- Brightspace Assignments
- In Brightspace Assignments, you can switch on the plagiarism scanner:
- Ouriginal (will be replaced in 2023)

You can contact [Teaching & Learning Support](#) if you need more information, or if you want to use other tools for assessment.

For more information on all centrally supported tools, including (peer) feedback tools, please see [this page](#).

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 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 document with the *teaching and examination regulations* (TER).

1.7.b 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 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 finetune 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 in Table 4. 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.

1.7.c 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

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>

1.8. 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, and your individual assessments meet the quality requirement of your assessment. Some of the requirements are explained in more detail in the following checklist:

1.7.d 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

Checklist 1: Summary of quality requirements for assessment

Quality requirement for assessment	Description
Validity	<p>Validity is also called <i>representativeness</i> (whether the assessment represents the content and level of the learning objectives). This implies the following:</p> <ul style="list-style-type: none"> - The tests <u>cover</u> the learning objectives and nothing else. - The tests are at the <u>level</u> of the learning objectives. - The assessment methods match the learning objectives. - The <u>weighting</u> 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. <p>Assessment <i>blueprints</i> (consistency check tables for assignments and assessment matrices for exams) visualize whether an individual assessment represents the learning objectives.</p>

Quality requirement for assessment	Description
Reliability	<p>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:</p> <p>Specificity implies that:</p> <ul style="list-style-type: none"> - Grades represent the level of mastery; - Only students who master the LOs to a desirable level can <i>pass</i> (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 <i>grade</i> 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. <p>Test-retest reliability implies that the same student should get the same score if they answer a question twice:</p> <ul style="list-style-type: none"> - 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 <i>difficulty</i> over the years; - <i>Enough questions</i> are asked to get a good sample. <p>Objectivity implies that the grade does not depend on the grader (rater), i.e. the <i>rater bias</i> is minimised.</p>
Transparency	<p>Making grading criteria and methods known and clear to students:</p> <ul style="list-style-type: none"> - 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	<p>Also referred to as 'usability'. This relates to the workload and availability of resources, for example:</p> <ul style="list-style-type: none"> - 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 lecturers and teaching assistants to prepare, provide feedback and grade the assessments?

Quality requirement for assessment	Description
Efficacy	<p>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:</p> <ul style="list-style-type: none"> - Is the assessment <i>authentic</i> (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 <i>feedback</i> effective for the student? - Do students get <i>feedback</i> on their performance on each learning objective <i>before</i> 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. This section will discuss grade calculation, and alterations that could be made after a test result analysis.

2.1. 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

In the Assessment Framework, this will be changed as follows:

9.5-10.0	Excellent	Uitmuntend
8.5-9.0	Very good	Zeer goed
7.5-8.0	Good	Goed
6.5-7.0	Very satisfactory	Ruim voldoende
6.0	Satisfactory	Voldoende
5.0-5.5	Almost satisfactory	Bijna voldoende
4.0-4.5	Unsatisfactory	Onvoldoende
3.0-3.5	Very unsatisfactory	Ruim onvoldoende
2.0-2.5	Poor	Slecht
1.0-1.5	Very poor	Zeer slecht

In addition, course results can have the following values:

NV	No show	<i>Niet verschenen</i>
	if a student registered for an assessment but did not show up / did not hand in their work.	
NVD	Did not pass	<i>Niet voldaan</i>
	if a student e.g. did not receive sufficiently high assessment results, or did not participate in some parts. Consequently, a final grade cannot be calculated.	
VR	Exemption	<i>Vrijstelling</i>
	if the Board of Examiners granted an exemption.	

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. Grade 1 is by Dutch convention the lowest grade that a student can obtain.

2.1.a What does a minimum pass grade imply?

Grade 6.0 (or 5.75 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 course examiner should determine what the minimum level is 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.

2.1.b (Dis)advantages of increasing test requirements to test LO achievement

What about having one exam per learning objective, and requiring a 6.0 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. Nobody is able 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 the number of tests and their accompanying minimum grades are increased, this increases 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 assessments, which would mean a lot of work for you. 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, prevent creating unnecessary assessment hurdles.

On the other hand, it is important to gain insight on which learning objectives are accomplished by the students, and which not. Your course is, after all, part of a larger programme 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 both 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).

2.1.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 objective 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.

2.2. Grade calculation

Now that we are clear on what a grade is, we will continue to how to calculate grades.

2.2.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: $Grade = 1 + 9 * \left(\frac{score}{\max(score)}\right)$
Dark blue triangles: $Grade = \max \left\{ 1; 10 * \left(\frac{score}{\max(score)}\right) \right\}$

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.

You should check if one of the above is mandatory / advised in the assessment policy of your faculty or programme. The first one is most commonly used.

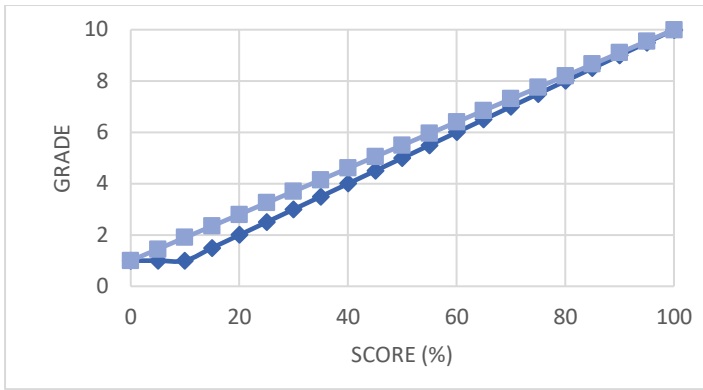


Figure 2. Two simple score-grade transformation. Horizontal axis: relative score (percentage). Vertical axis: grade. Light blue squares: 0 points lead to a 1, the grade increases after each point earned. Dark blue triangles: grade runs from 0 to 10 and rounded to 1 for grades smaller than 1.

The method you choose will determine the *cut-off score*: 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 light blue squares, grade 6.0 corresponds to collecting 55% of the points, while for the dark blue triangles, 60% of the points will assign the student grade 6.0.

The examiner will communicate the cut-off score or the score-grade formula on an exam's cover page or in the assessment instructions.

2.2.b Score-grade transformation and cut-off score for closed-ended questions

When calculating the grade for MCQs, you can 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.

You should check if the guessing correction is mandatory/advised in the assessment policy of your faculty or programme. 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 $\frac{1}{4} = 25\%$, 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 MCQs, the guess correction should be considered, such that the students will have no points (or a 1) whenever their score is equal or lower than the guessing correction.

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

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

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

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.

2.2.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. The cut-off score should reflect the minimum level that students should have reached in order to pass the course. 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. 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 split score-grade transformations with a cut-off score of 16 points (blue circles) and 32 points (orange squares) respectively. This representation is for closed-questions.

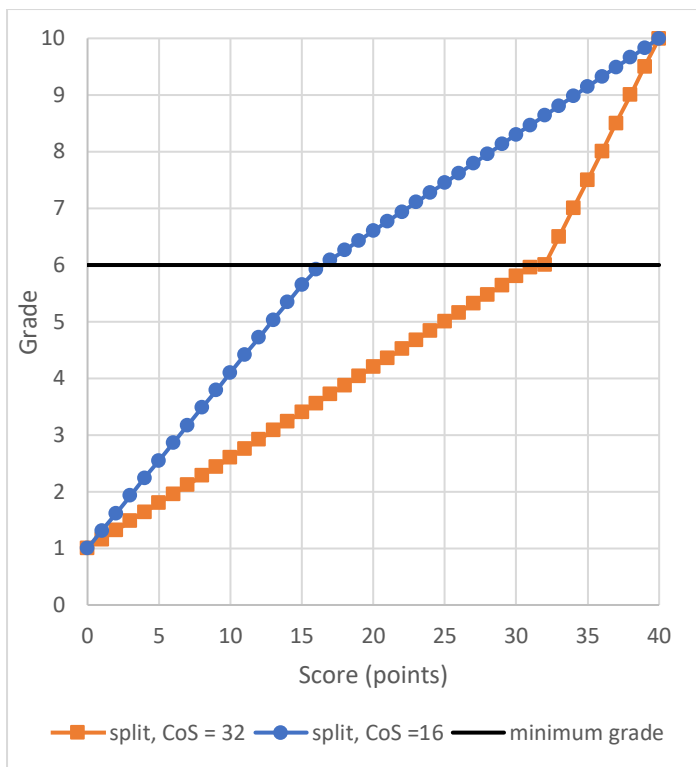


Figure 3 Score-grade transformations for two split transformations around cut-off scores of 16 points and 32 points. CoS = cut-off score. See running text for the formulas.

Used formulas for split transformations:

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

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 and *ms* the maximum score.

2.2.d Closed-ended questions, knowledge percentage and relation to cut-off score

In exams, the *knowledge percentage* is the percentage of questions that a student 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](#))?

2.2.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 lecturers, it is important to distinguish correctly and consistently 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 point, that would mean that a change of 1 point changes the grade by $(10 \text{ (maximum grade)} - 1 \text{ (minimum grade)}) / 50 \text{ (total points)} = 0.225$. This means that 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. $(10 - 6 \text{ (minimum pass grade)}) / (60\% \times 40) = 0.17$ points, and coarser for the fail grades, i.e. $(6 - 1) / (40\% \times 40) = 0.31$.

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). The gradient of the line changes at the cut-off score: 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.

2.2.f Exams with both open-ended and closed-ended questions

If your exam consists of 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 consider the guessing correction. After calculating both grades, you 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 them feedback on what type of questions they need to focus on most during their preparation for future assessments.

The reason why you need to calculate the grades separately, is so a guessing correction can 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 your students did not get any points for the MCQs (0 points) and full points for the open questions (60 points).

Firstly, consider the situation in which you 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 you 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 exam, 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		
		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 Table 6, 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.

2.3. Objectivity and reliability of grading

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 all human, it is nearly impossible for anybody not to occasionally make errors while 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.

2.3.A Inter-assessor reliability

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

- *Generosity errors*: 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.

2.3.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 and the measures that can be used to avoid them:

1. 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 whom does not know the students - evaluate the results.
 - Use an answer model or a rubric.
2. Contrast effect:
the assessor over- or underrates 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 than the rest.
3. Sequence effect:
there is a shift in standards, or the scoring criteria are redefined over time.
 - Use an answer model or a rubric.
 - Evaluate per question – not per student, and change the order of the students per question.

CHAPTER 3) ANALYSIS OF TEST RESULTS

When considering how well you and your students performed, you 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.

Test results analyses take place on three different levels (see Figure 4):

- 1) At test/assessment level
- 2) At item level
- 3) At answer level

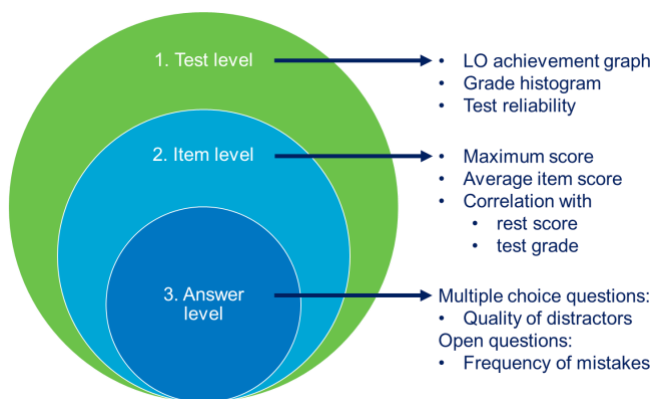


Figure 4. Three levels of test result analysis

A test result analysis will give insight into:

- 1) How well the students mastered the individual learning objectives of the course (test level, see **Error! Reference source not found.**)
- 2) The quality of the individual test questions or assignment criteria (item & answer level, 0)
- 3) Whether the answer model need to be revised (3.3)
- 4) The overall quality of the assessment (test level, 3.4)
- 5) Whether the grading needs to be revised (3.5)
- 6) How to adjust rubrics and criteria (3.6)

This chapter explains the steps that you can take to perform a test result analysis and to improve the grading, future assessments and future courses based on your findings.

Keep in mind that it is practically impossible to make flawless assessments (unless you had unlimited time). Therefore, be prepared to adjust the answer model or rubric grading after the test result analysis.

First, please take note of the following definitions:

- *Test*: any assessment, including projects, assignments, exams with open-ended questions and multiple-choice exams.
- *Grade*: the grade (usually on a scale from 1 to 10) that a student receives for the whole test
- *Score*: the number of points that a student obtained from this test, before it is transformed into a grade.
- *Item*: the smallest unit in a test. This can be a criterion or subcriterion for assignments/projects, or a subquestion or question for an exam.

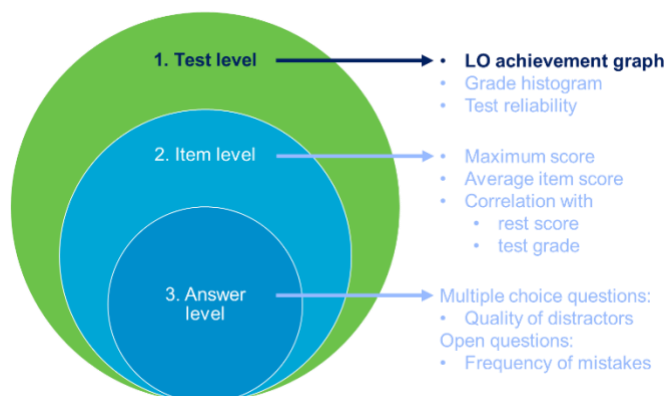
In case your exam is a digital exam or a paper-scan exam, digital exam tools do the test analysis result for you (in case of Ans and Brightspace Quizzes). Check the Teaching Support pages for an explanation and how to use this part of the assessment tool. More information on how to interpret test result analyses, see sections **Error! Reference source not found.**, 3.4.

If you are grading exams or projects/assignments with pen-and-paper, you will store the following data in a spreadsheet while (or after) scoring your students' work:

- Scores per item per each student
- Total scores per student
- Grades per student

For smaller datasets (few students (<20) or few items), you may not be able to draw strong conclusions from your data. However, you are encouraged to run a test result analysis to check if your experience during the course and grading matches the test result analysis.

3.1. Analysis of the achievement of 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 5.

Graphically summarizing the scores of your students per learning objective will make it 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 data points.

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. You can use any graph of your choice, as long as it summarises the distribution of the scores per learning objective.

Typically, problems in learning objective achievement are caused by a lack of practice at the level of the test (constructive alignment).

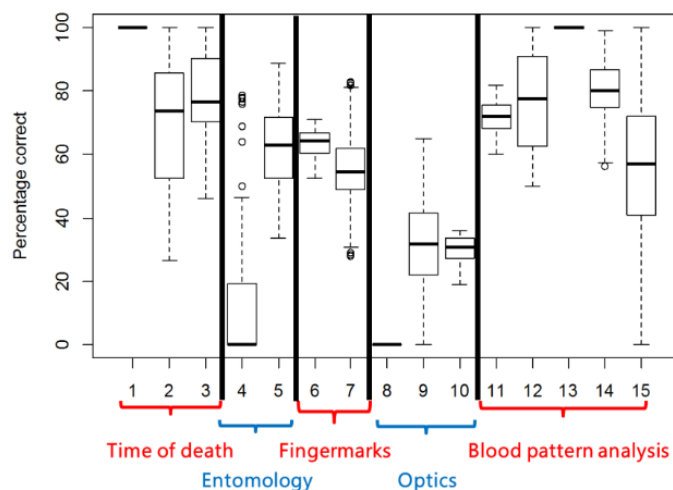
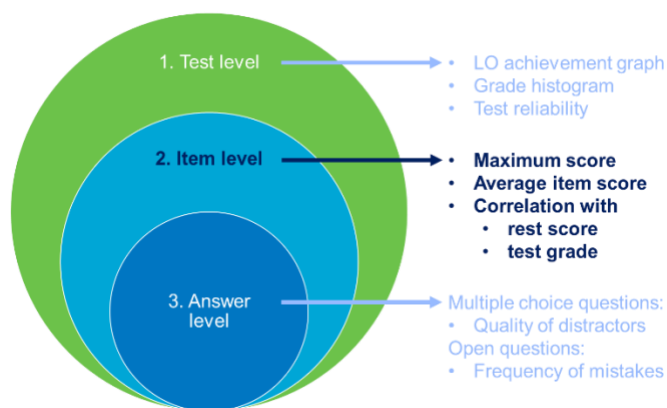


Figure 5. Example graph indicating test scores and LOs in a boxplot

3.2. Analysis of the quality of the test items and answers



In this section, you will learn to analyse the quality of the individual items ((sub)questions, or assessment (sub)criteria) with the outcome of the item-specific analyses. Use this to pick the most worrying items to check for errors or unclarity. This helps you to improve the scoring of these items for the sake of the students who just took the exam and will have a fairer grade, and for the sake of selecting how you are going to further improve next year's test.

Test result analyses result in a number of variables. The most useful variables to check which items are (most) worrisome are the following:

- Maximum score: Did at least a few students answer this individual question correctly or get the full score for this criterion?
- Average score: Is the average score very high or very low (and did you expect this)? I.e. was the question / criterion very easy or very hard?
- Correlation with the other scores: How did the good-performing students do on this question or criterion?

Use these variables to pick, for example, four items to study in detail.

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

3.2.a Maximum score achievement

The goal of your course is to facilitate your students to master the learning objectives, and the goal of the assessment is to measure whether you and they succeeded. For each individual item, you would expect that there are students who get full score (if there are a reasonable number of students). Therefore, you check for example the maximum score (*maxa* in the [TU Delft Excel](#)), expressed in points.

If no (or too little) students got the full score for an item, 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 (considering 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?

3.2.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.

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*. Confusingly, *p* is called the 'difficulty', although the higher the value of *p*, the 'easier' the item.

$$p = \frac{\text{Average score}}{\text{Maximum score}}$$

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*.

But how can you determine what *p*-value is okay? When designing an exam, you want to include questions that cover a wider range of difficulty, so that the test can distinguish between good and very good performing students, as well as between pass and fail students. Most important is to check whether the difficulty matches your expectations. 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 that distinguishes between pass and fail students is in the range between 0.4 and 0.6 (See 2.2.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 get 50% of the points on average. It would upset your students and depress yourself during the scoring process.

In case of MCQs, *p* is ideally halfway between the guessing score ($1 / (\text{number of options})$) and 1 (see Table 7). Some programs like Ans also calculate a *p* that is corrected for guessing (*p'*), meaning that a *p'* of 0 is defined as the guessing score.

Table 7. 'Ideal' *p*-values

Number of options	Guessing score	Ideal <i>p</i> -value	Ideal <i>p</i> -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 ($1 / \text{number of options}$, see Table 7), this might indeed have been caused by guessing, for example because the topic was not included in the course. If *p* is lower than the guessing score, there either is a misconception amongst students, or another option might be the correct answer instead.

Note: See 2.2.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.

3.2.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 three item discrimination coefficients: R_{it}, R_{ir} and R_{ig}. You can always use R_{ir}, but not always the other two.

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.

Terminology: 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, and ‘ig’ for item-grade. All three measure the correlation between the item score and a measure of ‘true student score’:

If available, use the R_{ir} and ignore the R_{it}. 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 or if not all questions have the same weight/amount of points. In that case, the R_{ir} score is more reliable (less biased by e.g. outliers). In other cases, the difference between R_{it} and R_{ir} will be low.

In some projects/assignments, lecturers do not calculate the grade directly from the criteria scores. Instead, they determine the students’ grades separately and use the rubric/criteria scores to explain the grade. In these cases, the item-grade correlation helps to determine the (unconscious) importance of the individual criteria for determining the grade.

The R_{ir} of items is calculated using the following formula:

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

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 item score of student i on item j , and μ_j the mean score on item j .

The R_{ir} of subquestion j is calculated using the following formula:

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

Where $\tilde{\mu}_j$ is the mean test score calculated from all subquestion scores minus the score from item j . The R_{ir} and R_{it}-values are always between -1 and +1². These values can be interpreted as follows in case of closed-ended questions:

Ideal values: Items with a R_{it}/R_{ir} of at least 0.20 (see Table 8) are considered sufficiently distinguishing between poor and good performers. 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 (see 3.6 on when R_{ir} values are too high and might require action). It is wise to always look at the lowest R_{ir}-values of a test.

Table 8.
Interpretations of R_{ir} and R_{it} 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.

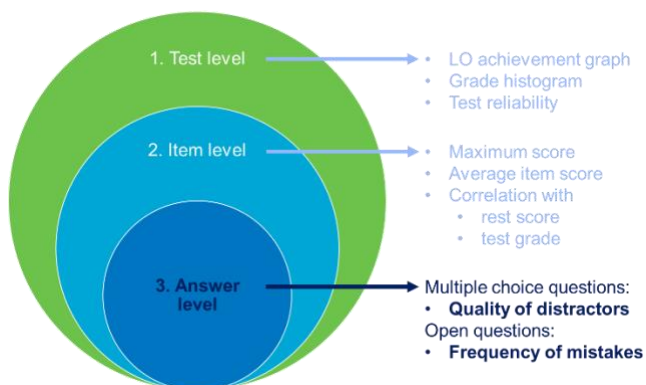
² R_{ir} squared equals the percentage of variance in the final grade that is explained by the score for the item. So if R_{ir} of question 4b equals 0.5, it indicates that 25%

of the variance of the final score (i.e. the grade) can be explained by the score of question 4b, under the assumption of a linear relation.

Negative values: In case R_{ir} is quite negative (i.e. not ~ 0), 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 R_{ir} 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 R_{ir} 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.

3.2.d Quality of 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 -values (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'.

Underlying issue: 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 a 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.

3.2.e Quality of distractors MCQs (average total score per option)

Use the average total (rest) score of students of the options to check whether the better performing students chose the correct answer, and not a specific distractor. You want this value to be higher for the correct answer compared to each distractor's value. This check is possible in the TU Delft Excel for mc exams.

Underlying issue: There might be a distractor that lures otherwise good scoring students into overthinking a question. Check whether these possible trick distractors are also (partially) correct and consider giving students full or partial points. You might also have made a mistake in the answer key (happens to the best of us).

3.2.f Frequency of mistakes (open questions)

If you are grading open-ended questions using Ans, the tool keeps track of how often students get a step in a calculation right (or wrong), and how often they make specific mistakes. This gives you a good indication of what your students do and do not master, and if specific issues might be related to unclarity in the test assignment, in the answer model, in the course material, or during lectures/tutorials.

3.2.g Finding the most worrying items

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

- (almost) no students got the maximum score,
- R_{ir} s are negative or relatively low
- p -values are low
- a -values are high (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.

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.

3.3. How to adjust answer models based on test result analysis

The last section 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. This section discusses 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, you will find out how you could change the grading.

3.3.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 R_{i-s} 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.

3.3.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:

Indicator for worrying item	Implication
1) $Max_a < max$	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) $R_{i-r} < 0$ (e.g. -0.2)	<p>Good students performed not good on this question, and/or not-so-good performing students performed good on this question. This is always problematic.</p> <p>In case p is small, this indicates that the few students who answered the question correct, were the bad-performing students.</p> <p>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?</p>
3) $R_{i-r} \sim 0.0$ (<0.2) or for open questions: the lowest R_{i-r} -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 than the correct answer. Especially if the R_{i-r} 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 R_{i-r} 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.

- 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.

3.3.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 (considering 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 R_{ir} S:

- 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 R_{ir} s might indicate that too many items are measuring the

same thing. You might consider calculating the correlation between all individual items to check whether this is true.

Low R_{ir} ?

- 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'.

3.3.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 R_{ir} -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.

3.4. 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 you not so sure, due to measurement errors? And does the student with a 10.0 really master all learning objectives? One way to estimate the measurement error is to calculate the score reliability (reliability coefficient), like Cronbach's alpha.

Assumptions of reliability: All reliability coefficients assume that the test intends to measure one single thing, namely how well as student masters the course. It also assumes that each student should perform more or less equally well on all test items, considering the fact that your job as teachers is to help student master *all* learning objectives of a course. If your students participate in all learning activities of your constructively aligned courses, you would find it 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 (i.e. *internal consistency of the test*). 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** 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)	$\alpha \geq 0.8$
Medium/low stake assessment (e.g. 50% of final grade):	$\alpha \geq 0.7$
Formative assessment (e.g. 0% of final grade)	$\alpha \geq 0.6$

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 3.2.c on page 28). This is frequently encountered in multidisciplinary master courses.
- **Test heterogeneity:** If the items represent very different topics or skills, this will lead to a lower reliability coefficient.
- **Mostly low or high scoring items:** the reliability coefficient will be lower if there are mostly items that result either in a low score in most students, or a high score in most students. Consider including items of average difficulty.
- **Little difference between student levels:** the reliability coefficient will be lower if students are at more or less the same level.
- **Low item correlation:** lower quality items (with higher R_{ir}) decrease reliability of the entire test (see 0 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 items, σ_x^2 the variance in the total scores of all students, i.e.:

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

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 item scores σ_j^2 is calculated equivalently:

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

with s_i the sub-question score of student i on sub-question j , and μ_j the mean score on sub-question 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 sub-question. The reliability coefficient can have a value between 0 (unreliable) and 1 (reliable). In very rare cases it can be negative. In a reliable test, the variance in the final scores of the students (σ_x^2) is much larger than the sum of variances in the sub-question scores (σ_j^2).

3.4.a Confidence interval of grades (standard error of measurement)

The meaning of reliability will be illustrated by discussing how you can use Cronbach's alpha to calculate the *measurement error* that was introduced by chance. 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 (unbiased) 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 consists of the true score plus the measurement error, either systematic or accidental. To ensure that grades are correct and, more specifically, that students correctly pass or fail the course, it is important that the error of measurement is as small as possible.

You can calculate the measurement error as follows: First, you calculate the **Standard Error of Measurement (SEM)** from Cronbach's alpha or KR-20:

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

in which x is the achieved test score, SD is the standard deviation and α is the reliability coefficient (Cronbach's alpha or KR-20).

From here, you can calculate the 68% (most common) or 95% **confidence intervals** in which the 'true score' of the student lies:

Table 9.
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]

Meaning of confidence interval: The confidence interval indicates that if the student would repeat the test for an infinite times in the same circumstances, the average grade (and hence the *true grade*) would 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.

Example:

- a student scores 26 out of 50 points
- the cut-off score is 28
i.e. grade of 6.0 (grade = 1 + 9*score/50)
- SEM is 5 points
- the 68% confidence interval is 21 to 31 in points
i.e. a grade between 4.8 and 6.6
- The student will get a 5.7 for the test, which is rounded to a 5.5 (fail) if the course consists of one test.

This means that the student has failed, but maybe should have passed based on his true score (actual capacity) and wasn't able to because of the either systematic or accidental measurement error.

The 95% confidence interval is even wider:

- the 95% confidence interval is 16 to 36 points
i.e. a grade between 3.9 and 7.5

Use of confidence intervals: The confidence intervals can be used to determine which students' work might benefit from a second reviewer to (independently if feasible) rescore these students' work. This could be the case for students whose confidence intervals contain the cut-off score.

Consequence:

The uncertainty of grades is a reason to allow for compensation between partial grades within a course, and in some cases even between courses. The latter is up to the Board of Examiners to decide.

Example: The Board of Examiners 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 build on the ones in the first course.

3.4.b Frequency distribution of grades

You can represent a **frequency distribution** of the grades in a histogram, or a **cumulative percentage**, like in Figure 6. Use this to decide whether or not to increase the grades, based on whether your experience during the course is that is or is not an accurate reflection of the level of the students. This might depend on the percentage of students that passed the course.

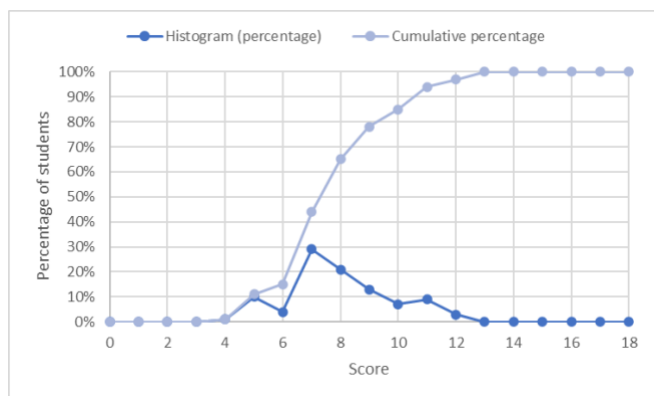


Figure 6. Example of frequency distribution of grades (histogram and cumulative) for a test with maximum 18 points.

In Figure 6, if the cut-off score is set to 10 points ($grade = 1 + 9 * score / 18$), only 22% of the students will pass. You could use the histogram to determine a new cut-off score.

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 56% 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** (see Figure 7), most students have a relatively low score, meaning the test was too difficult for a large group of students. When you have a **ceiling effect** (see Figure 8), most students have a relatively high score, meaning that the test was too easy for a large group of students. In case of courses that have a 'steep learning slope' (i.e. require a relatively large portion of the course to reach a basic level), you might see a combination (camel) effect with two bumps (one at a very low grade for the students who did not reach

the basic level, and one centred around e.g. 7.5). You are the expert who knows what happened during the course and are the expert at explaining your grade distribution.

Examples of both are shown below:

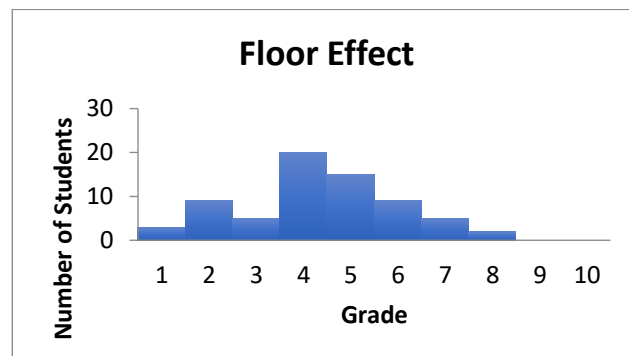


Figure 7. Grade histogram demonstrating the floor effect

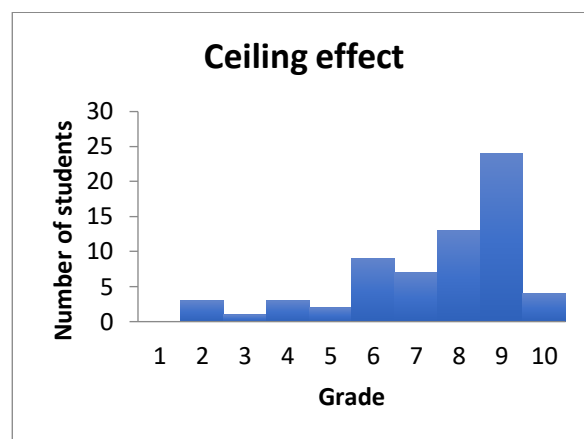


Figure 8. Grade histogram demonstrating the ceiling effect

3.5. Adjusting the grades

3.5.a 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 simplest one is to simply add a constant number to the grade. Another way is the *Cohen-Schotanus adjustment*. This one is described below.

3.5.b Cohen-Schotanus' adjustment of score-grade transformation

Cohen-Schotanus (University of Groningen, Medical Faculty) explains that because lecturers could (and often do) make mistakes with their exams (and courses), it is possible to underestimate students' abilities. In short, her method assumes that the top 5% of the students is supposed to get a 10. Therefore, it calculates the *average score of the top 5% students* and assigns them a 10. This method uses a **knowledge percentage** 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.

3.5.c 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.

3.5.d 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?

3.6. How to use correlation table to adjust criteria or rubrics

In the previous sections, some ideas on how to adjust/create a rubric based on the test result analysis were already described. In this section describes how you can use the correlation **between** criteria to update your assessment criteria and/or rubric (see Table 10).

What does it mean? The correlation between criteria indicates if two criteria increase and decrease together. If they do, the value is positive (between 0 and 1), and if they 'anti-correlate', the value is negative (between -1 and 0). This happens if students who do relatively well on one criterion, actually do worse on another criterion.

What can I do with this information?

If some criteria have a **large positive correlation**, this implies that they are scored pretty similarly.

- There may be too many criteria, causing the assessors to become 'lazy'. Consider **combining** these two criteria, but only if it makes sense to combine them.
- The criteria may not be distinctive enough to assessors, although you think that there is a clear and meaningful distinction. Consider rephrasing the criteria, their description or their rubric descriptors, to clarify this.
- Example in Table 8: 'correctness' and 'design' have a large correlation. Assessors might think that they are the same, judging by their names. You might consider combining them, or if you find their difference very important in this course, find a way to clarify their distinction to the assessors.




If some criteria have a 'significant' (e.g. smaller than -0.1) **negative correlation**, this implies that students

who do well on one criterion, do relatively bad on the other one.

- Either or both of the criteria may not be trained during the course (i.e. there is a constructive alignment issue). Consider training students on this criterion and giving them feedback, or taking out the criterion.
- Example in Table 8: Students who do well on the summary, do not do so well on the presentation, and vice versa. Maybe the lecturer did not train students or give them feedback on the presentation or summary? Or gave half the students feedback on the summary, and the other half on the presentation because there was not enough time for them to hand in drafts for both?

If two criteria have **no correlation** (e.g. between -0.1 and 0.1), this implies the criteria are independent. That is actually not per se a bad thing, since you want the criteria to measure different things. Why? If they measure the same thing, it can be a waste of time. However, you would expect some positive correlation, because the assumption is that students learn all criteria in equal proportions (i.e. they all have the same 'easiest' and 'hardest' learning objective). Therefore, you may want to double check for pairs of criteria that you expected to correlate and did not correlate. If you find any, check if you provided sufficient training and feedback.

Table 10. Correlation between criteria in a project

Legend:  = positive correlation
 = no correlation
 = negative correlation

CRITERIA	Grade	Speed	Fuel consumption	Design	Production costs	Aerodynamics	Summary	Correctness	Completeness	Presentation	Contribution
Speed	0.34	1.00									
Fuel consumption	0.40	0.12	1.00								
Design	0.69	0.56	0.43	1.00							
Production costs	0.52	0.15	0.03	0.60	1.00						
Aerodynamics	0.50	0.58	0.24	0.59	0.28	1.00					
Summary	0.45	0.26	0.34	0.48	0.37	0.31	1.00				
Correctness	0.72	0.04	0.30	0.59	0.59	0.38	0.31	1.00			
Completeness	0.21	0.03	0.04	0.24	0.34	-0.05	0.06	0.18	1.00		
Presentation	0.29	0.09	0.28	0.13	-0.07	-0.05	-0.12	0.13	0.46	1.00	
Contribution	0.66	0.32	0.25	0.49	0.50	0.17	0.33	0.51	0.27	0.21	1.00

CHAPTER 4) CREATING AND IMPROVING PROJECTS/ASSIGNMENTS



Figure 9. Assessment cycle for courses with projects / assignments

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 5)) and assignments, the process is very much alike:

Table 11. 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 10).

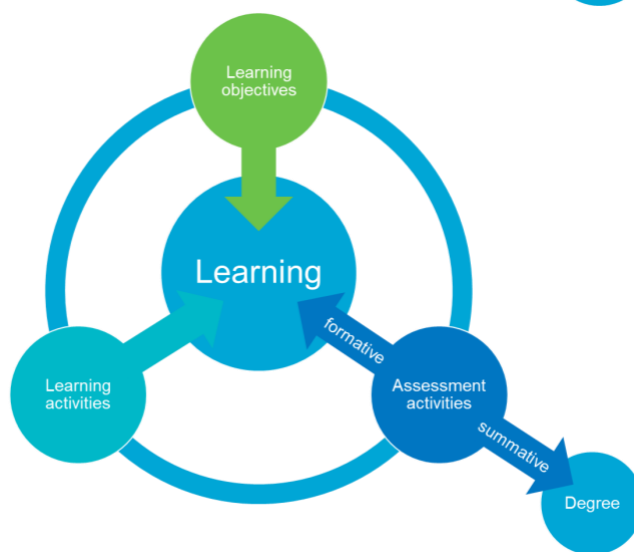
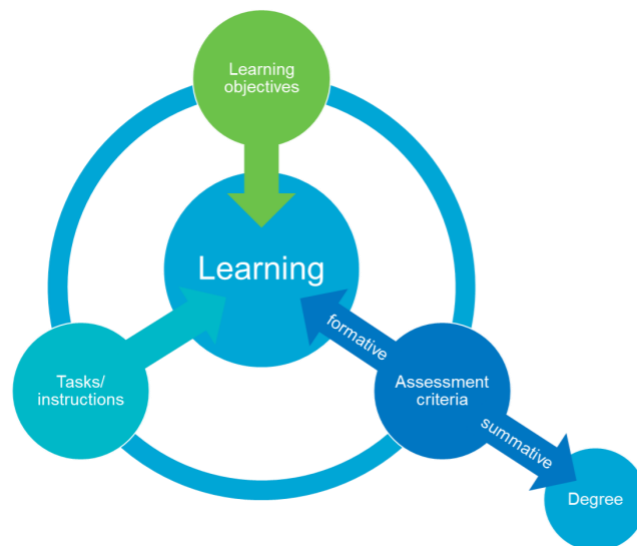


Figure 10. Constructive alignment triangle for a course (top) and for an assignment (bottom)

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.

4.1. 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 12.
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, 60%)	Contribution (individual, 15%)	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				- Contribution to group process (5%) - Contribution to product (5%) - Reflection on group process (individual, 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).

4.2. 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 4.4.b 'Checklist for assignments', you will find the checklist that may help you to formulate your assignment.

4.3. Assessing assignments: rubrics and grading instructions

4.3.a Using rubrics

In an answer model for assignments, you can either use a simple scoring guide rubric or a rubric. 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** 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 13).

Table 13.

Rubric for grading the group part of the report of the bridge designing project from the consistency check table in Table 12

LEVEL CRITERIA (%)	Excellent (10)	Sufficient (6)	Insufficient (2)	Score
Exploration (25%)	At least 5 innovative and plausible options are detailed 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 & decisions (33%)	The decision is based on a trade-off between all quality criteria and is based on valid arguments.	The decision is based on a trade-off between most quality criteria and is the argumentation is valid.	The decision is not based on the quality criteria ----- or ---- the argumentation is missing.	
Drawings (17%)	The drawings provide a excellent overview of the structure as well as all essential structural details.	The drawings provide a rough overview of the structure and 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.	
<p>Knock-out Criteria -Grammatical and spelling errors do not severely hinder readability -Use of required report structure</p>				
<p>Grade Calculation = ,25*exploration + .33* Considerations & decisions +.17*Drawings + .25*Calculations This grade counts for 60% of the final grade</p>				

4.3.b Grading instructions

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 **Error! Reference source not found.** in section **Error! Reference source not found.** on page 51.

4.4. 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 for 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.

4.4.a Checklist for consistency check tables

Checklist 2. Checklist for consistency check tables

Checklist for consistency check table
<ul style="list-style-type: none"> <input type="checkbox"/> Are the criteria in the rubric are the same as in the consistency check table? (<i>validity, alignment</i>) <input type="checkbox"/> Are the criteria names are short, descriptive, specific and clear? (<i>reliability, transparency</i>) <input type="checkbox"/> Do students get (peer) feedback on all criteria first before being evaluated for a grade on these criteria? (<i>effectivity</i>) <input type="checkbox"/> Is each learning objective fully covered by its criteria? (<i>validity</i>) <input type="checkbox"/> Are the criterion weightings are representative of the importance of the learning objectives?³ (<i>validity</i>) <input type="checkbox"/> Are all criteria that do not match a learning objectives knock-out criteria? (i.e. prerequisite to receive feedback or grading; <i>validity</i>) <input type="checkbox"/> Are the criteria unique? (no overlap between criteria) (<i>reliability</i>)

4.4.b Checklist for assignments

Checklist 3. Checklist for assignment description

Checklist for assignment description
<ul style="list-style-type: none"> <input type="checkbox"/> Are the students addressed directly? ('you will' instead of 'the students will') (<i>effectivity</i>) <input type="checkbox"/> Is the lay-out clear? (e.g. use of bullets for steps, highlighting what is important) (<i>effectivity, transparency</i>) <input type="checkbox"/> Are resources provided (literature, formats, example code, etc.), if finding/creating them is not part of the learning objectives? (<i>validity, effectivity, practicability</i>) <input type="checkbox"/> Is the assignment written clearly and concisely. (<i>reliability</i>) <input type="checkbox"/> Is all terminology likely to be known to all students? (e.g. no regional/national 'general knowledge') (<i>reliability</i>) <input type="checkbox"/> Is the assignment aligned with the learning objectives? (<i>validity</i>) <input type="checkbox"/> Is there enough time to complete the assignment? (<i>practicability</i>) <input type="checkbox"/> Will the assignment lead to a product that will demonstrate the level of mastering the criteria? (<i>validity</i>) <input type="checkbox"/> Does the assignment description contain each of the following elements? (<i>effectivity</i>): <input type="checkbox"/> introduction: stating the relevance of the assignment. <input type="checkbox"/> learning objectives: stating what the student will learn. <input type="checkbox"/> instructions: explaining the activities that need to be undertaken. <input type="checkbox"/> product: describing what the concrete result are. <input type="checkbox"/> feedback/evaluation: criteria for assessment, <u>and</u> when and how feedback will be given.

tion 4b.

³ Henk van Berkel, *Zicht op toetsen*, 1999, Van Gorcum, pp 152-153.

³ 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.

4.4.c Checklist for rubrics

Checklist 4. Checklist for rubrics

Checklist for grade
<input type="checkbox"/> Is it clear what the weightings of the criteria are? <input type="checkbox"/> Is it clear how the grade is derived? <input type="checkbox"/> Does performance at the minimum level of a pass leads to a pass grade ? <input type="checkbox"/> Is it possible to get a 10 , judging by the criteria descriptors?
Checklist for descriptors
<input type="checkbox"/> Is it feasible to get a 10 , judging by the descriptors of the highest levels ? <input type="checkbox"/> Are the descriptors objectively formulated? (no 'just sufficient' 'excellent') <input type="checkbox"/> Are the descriptors specific and clear ? <input type="checkbox"/> Are the descriptors of each criterion unique ? (no overlap between descriptors of adjacent levels).
Checklist for usability
<input type="checkbox"/> Does the rubric give a good overview at first glance ? (not too many rows or columns) <input type="checkbox"/> Does the rubric fit on one A4? <input type="checkbox"/> Is the lay-out clear? <input type="checkbox"/> Is the amount of details suitable? (not too detailed / no information that belongs in a course book). <input type="checkbox"/> Is there space for specific (individual) feedback ?

4.5. 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 to answer:

- Do you assess the students on soft skills like 'group skills'?
- Do you 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 Scorpion. 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).

CHAPTER 5) CREATING AND IMPROVING EXAMS

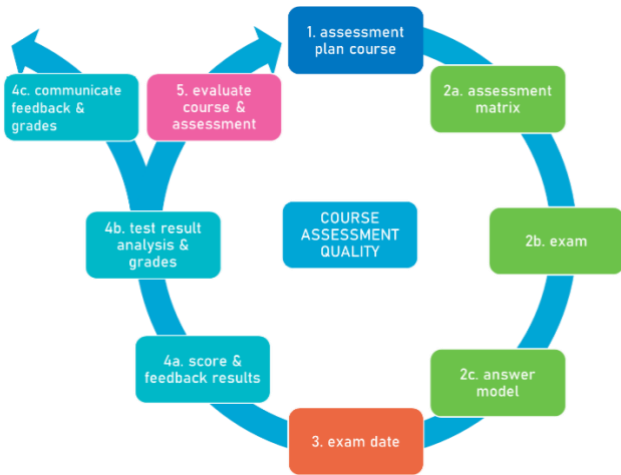


Figure 11. Assessment cycle for courses with exams

Designing an assessment has four stages:

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

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

Table 14. Comparison of assignments and exams

	Assignment	Exam
1. Blue print	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, including knock-out criteria	Exam, including front page

3. Grading guide	Rubric (and/or assessment sheet): - instruction for graders - knock-out criteria	Answer model: - model answers - points to be awarded in each situation - instruction for graders
4. Feedback from stakeholders	Experienced and new assessors, students	Experienced and new assessors

5.1. Exam blue print: assessment matrix

5.1.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). 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 course you teach has a set of learning objectives or learning outcomes. This manual assumes is that you use Bloom's taxonomy to define those learning objectives. If your programme uses another taxonomy, that is fine, too, of course. 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.

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.

Figure 12: Example set of learning objectives for a course

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),

Table 15.

Example assessment matrix for an existing exam based on the learning objectives listed previously. Q = (sub)question number, P = points per (sub)question. Dark blue cells indicate the level that the learning objectives in the left column describe.

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	3							10
	1b	4							
	3a	3							
2	1c	5							10
	4	5							
3			1e	3					8
			2a	5					
4			1e	5	2c	5	3	5	15
			2b	5					
5a			1d	7	1e	5			17
					4	5			
5b			3	5	3	10			20
6							3	10	20
							4	10	

Total	20	30	25	25	100
--------------	----	----	----	----	-----

5.1.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.

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 course need to be included. If the course 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.

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.

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.

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 adjust the cells, until you are happy.

If you are designing a new exam, for a new or redesigned course, 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 16):

Table 16. Example assessment matrix for a new exam. The dark blue cells indicate the level of Bloom that the learning objectives in the left column describe.

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.

5.1.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 be 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.

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. 'hours' means 28 hours * the number of ECTS 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.

5.1.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 too coarse? If not, add more questions in order to make the steps smaller.

5.1.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 are:

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.

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.

5.2. Assessing exams: answer model and grading instructions

5.2.a 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)*.

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 **Error! Reference source not found.**, 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 5.
Checklists for answer models

Include the correct / an ideal answer
<input type="checkbox"/> Are all possible answers included? <input type="checkbox"/> Are guidelines included on how many of these possible answers are required to earn points? <input type="checkbox"/> Are instructions included on the process to handle correct student answers that are not (yet) included in the answer model?
Include the maximum number of points
<input type="checkbox"/> Are the max points included both for main and subquestions? <input type="checkbox"/> Are the max points reasonable for the required amount of student work?
Description of how divergent answers are marked
<input type="checkbox"/> Is it clear which answers are considered fully/half/not correct? <input type="checkbox"/> Is it clear how many points the various half-correct questions will receive?
Be clear on how interrelated subquestions are marked
<input type="checkbox"/> Is it clear for assessors how points can be earned for interrelated subquestions? If the first subquestion is incorrect, can students still earn points for follow-up questions based on the incorrect value?

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.

5.2.b 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.

5.3. Checklist for exams

The most important hint is to write the exam questions together with the answer model, and use a colleague or other stakeholder to review them. Let your colleague check 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 colleagues.

5.3.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.

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

5.3.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 7.
Checklist for validity, reliability and transparency

Test only one learning objective at a time (validity)
<input type="checkbox"/> Do not try to cover more than one learning objective in the same question.
Relevance of each question (validity)
<input type="checkbox"/> Is it clear what knowledge or skill is being tested? <input type="checkbox"/> Is this knowledge or skill absolutely necessary in order to answer the question? <input type="checkbox"/> Is the answer model in line with what the test questions ask?
Language (reliability)
<input type="checkbox"/> Are there any spelling errors or typos? <input type="checkbox"/> Is the question unambiguous and is it clear what is being asked? <input type="checkbox"/> Have double negatives been avoided? Is the question concisely formulated?
Presentation
<input type="checkbox"/> Is the layout clear? <input type="checkbox"/> Are the figures clear?
Transparency of grading
<input type="checkbox"/> Before taking the test/assignment , do students know ahead of time what will be on the test both in structure and in content? <input type="checkbox"/> Before taking the test/assignment , did your students get experience with the types of questions with which you will be testing? <input type="checkbox"/> 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. <input type="checkbox"/> 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?

5.3.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 8.
Checklist for closed-ended test questions

Dos
<input type="checkbox"/> Do all questions end in a question mark? Students should be able to answer the question without looking at the answer ⁴ . <input type="checkbox"/> Do all distractors seem just as plausible as the correct answer? <input type="checkbox"/> Are all options are roughly of the same length ? <input type="checkbox"/> Are the right answers distributed randomly over A, B, C, D, etc.?
Don'ts
<input type="checkbox"/> Does no question inadvertently provides the answer to another question? <input type="checkbox"/> Are there no grammatical clues to indicate the right answer? <input type="checkbox"/> Are there no 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.

5.3.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 9.
Checklist for open-ended test questions

Item	Details to include
Use a 3-part Structure	<ul style="list-style-type: none"> <input type="checkbox"/> Context (optional) <input type="checkbox"/> Question (assignment) <input type="checkbox"/> Directions for answering, for example, 'Motivate your answer, showing which formulas you used. Write no more than 3 sentences'.
Be specific	<ul style="list-style-type: none"> <input type="checkbox"/> Use imperative sentences ("List three characteristics of X" rather than "What are the characteristics of X"). <input type="checkbox"/> Specify what you expect in the answer (e.g. "List the <u>three</u> characteristics of X"). <input type="checkbox"/> Avoid "anything goes" formulations such as "What do you think..."
Context and question	<ul style="list-style-type: none"> <input type="checkbox"/> Make sure the context is relevant for the question. If not, delete it.

Item	Details to include
should be linked	<ul style="list-style-type: none"> <input type="checkbox"/> 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	<ul style="list-style-type: none"> <input type="checkbox"/> For example, between old and new questions
Interrelated subquestions	<ul style="list-style-type: none"> <input type="checkbox"/> Can a student continue calculating with an imagined set of numbers if a first subquestion was answered incorrect? <input type="checkbox"/> If so, are students instructed on what value(s) to use?

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.

TABLES OF REFERENCE

Table of tables

Table 1. Assessment plan characteristics, divided into three assessment plan analyses.	5
Table 2. Example assessment plan	7
Table 3: Categories of learning outcomes (Nightingale et al, 1996)	10
Table 4. Overview of assessment related subjects (first column) that are covered in the Teaching and Examination Regulations (TER, second column)	15
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)	15
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	24
Table 7. 'Ideal' p-values	28
Table 8. Interpretations of R_{ir} and R_{it} values	29
Table 9. Confidence intervals of a test score, based upon the standard error of measurement (SEM)	34
Table 10. Correlation between criteria in a project	37
Table 11. Comparing design process for assignments and exams	38
Table 12. Example consistency check table for an imaginary 1st year bachelor project in which students have to design a foot-bridge.	39
Table 13. Rubric for grading the group part of the report of the bridge designing project from the consistency check table in Table 12.....	41
Table 14. Comparison of assignments and exams	44
Table 15. Example assessment matrix for an existing exam based on the learning objectives listed previously. Q = (sub)question number, P = points per (sub)question. Dark blue cells indicate the level that the learning objectives in the left column describe.	45
Table 16. Example assessment matrix for a new exam. The dark blue cells indicate the level of Bloom that the learning objectives in the left column describe.	47

Table of figures

Figure 1. Assessment cycle	4
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.	21
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.....	21
Figure 4. Three levels of test result analysis	26
Figure 5. Example graph indicating test scores and LOs in a boxplot	27
Figure 6. Example of frequency distribution of grades (histogram and cumulative) for a test with maximum 18 points.	35
Figure 7. Grade histogram demonstrating the floor effect.....	35
Figure 8. Grade histogram demonstrating the ceiling effect.....	35

Figure 9. Assessment cycle for courses with projects / assignments	38
Figure 10. Constructive alignment triangle for a course (top) and for an assignment (bottom)	38
Figure 11. Assessment cycle for courses with exams	44
Figure 12: Example set of learning objectives for a course.....	45

Table of checklists

Checklist 1: Summary of quality requirements for assessment.....	16
Checklist 2. Checklist for consistency check tables	42
Checklist 3. Checklist for assignment description	42
Checklist 4. Checklist for rubrics.....	43
Checklist 5. Checklist for exam cover pages	51
Checklist 6. Checklist for validity, reliability and transparency	52
Checklist 7. Checklist for closed-ended test questions	52
Checklist 8. Checklist for open-ended test questions.....	53
Checklist 9. Checklists for answer models	51

References

- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher education*, 32(3), 347-364.
- Cauley, K., & McMillan, J. (2010). Formative Assessment Techniques to Support Student Motivation and Achievement. *Learning 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/ocslid/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/ocslid/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.
- William, 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.

