Step-by-Step guide for examination

Target group: teachers. Revised version of the original Step-by-step guide for CW/CS + PSY, University of Twente. Reconstructed October 2015 for the faculty BMS. English version: Nov. 2016. Revised 2017/2024. Support editing: CELT (W.D.J. Vlas)
The guide was especially drawn up to support teachers who will be asked to submit information for an assessment screening. For the sake of quality assurance, the management and the Examination Board of a programme might want to get an impression of the quality of the assessments within their programme. Each year a module or courses of a programme can be selected and the responsible examiner(s) will be asked to supply needed information for screening. This guide helps examiners to understand what kind of information is needed and to supply this information.
Although the guide was designed for a special purpose, it gives a nice overview of the whole assessment process and can be used by any teacher to support the assessment design, grading and evaluation process.

Introduction

What are the important basic principles and focus areas for examination and assessment? If we are to ensure proper assessment of our students, three quality criteria are key: validity, reliability and transparency. How can teachers meet these criteria?
This document comprises a supportive step-by-step guide for those tasked with designing assessments. The subsequent explanatory supplement section provides further assistance on some of the topics of the Step-by-step guide. Occasional reference is made to sources of information located elsewhere.
You can find the Step-by-Step guide here: **Manuals | Assessment support** .

Since Oct. 2024 a new Testing & Assessment Toolbox is made available online. Still under construction, but a lot of information regarding assessment can be found here: Utwente: Testing & Assessment.

If you have any questions or require any sort of support with designing, grading or evaluating your assessment(s), you can get support of the CELT educational advisors for your faculty.
See the CELT-website: https://www.utwente.nl/en/ces/celt/who-we-are/

Tip: For those who are proficient in Dutch: this book provides a lot of useful information: Toetsen-in-het-Hoger-Onderwijs-van-Berkel-Bax-Joosten-ten-Brinke.pdf (3rd edition, 2014, published by Bohn, Stafleu en Van Loghum).
For more helpful links and information, please visit: Utwente: Testing & Assessment (English)
Other useful resources:

- Assess Teaching and learning Carnegie Mellon University, Eberly Center.
- Tip sheets: Assessment & Feedback, University of Waterloo, Centre for Teaching Excellence.

DEFINITIONS

To prevent misunderstandings about terms we use in this guide, we provide some definitions and explanation about the way we use terms like “test”, “exam”, “to asses”. The used terms have not a standard definition in general and the use of the terms for specific purposes is kind of ambiguous. In literature or in educational settings the terms might be used in a slightly different way.

USED TERMS	USED FOR
Examination Assessment	used in a generic, broad sense. Examination: the act of examining something; to test a candidate's knowledge, skills and attitude. Assessment refers to a wide variety of methods or tools that can be used to evaluate and measure whether students have achieved the learning objectives and acquired the required knowledge, skills and attitude. Methods like: written questions, oral questions, practical tasks, skill performance etc. In literature the term “exam” is also often used in a general way.
Test / testing	used more specific for written tests with open or closed (multiple choice) questions. Testing: the act of giving students a test (as by questions) to determine what they know and have learned.
Assessment method	refers to a method or format chosen for examination, like for instance a written test, a poster presentation or observed lab work. An examiner can choose from a variety of methods in line with his/her learning objectives and the framework conditions as stated by the educational programme and EER.
Assessment scheme / plan for the course, subject or module	a scheme or plan to show how examination takes place for a (Bachelor) module or a unit or (Master) course. It shows which methods are used, the learning goals that are examined by each of the methods, the way each method contributes to the overall grade for the module, unit or course (most of the time presented in percentages). It mentions also special conditions that may apply (e.g. a minimum grade or attendance requirement).
Test specification table (or matrix)	especially used for written tests. It shows how the questions are related to the learning objectives. It also indicates the level of the questions (knowledge, insight, apply etc. ; mostly the taxonomy of Bloom is used for this) and the attributed weight.
Grade / grading	number or symbol or term indicating a student's level of accomplishment at the end of a course, unit, module. In the Dutch system mostly a number from 1-10 or pass/fail is used as indication. Grading = to determine the quality of academic work. NB. In literature or conversations, the term “mark” is used often in the same way and in a general sense. Like: to mark an essay test or paper
to assess	to estimate or judge the quality of students’ work and assign a grade according to one's evaluation. The assessment is based on, for example, the number of correctly answered questions or based on criteria and standards when assessing assignments.
Assignment	a task or academic piece of work that has been assigned to students. It will show in the end whether they have mastered the required knowledge, skills, attitude, as stated in the learning objectives. If used in a summative way, students get a grade or pass/fail (or equivalent) for the results. If used in a formative way, students get feedback, without a grade. They can use the feedback as information to check and improve their work or learning process.

STEP-BY-STEP GUIDE FOR EXAMINATION AND ASSESSMENT

USE THE BLUE HYPERLINKS TO NAVIGATE TO MORE INFORMATION QUICKLY

Step 1) Learning objectives and choice of assessment method(s)

Learning objectives

(Re)formulate clear learning objectives for the subject component under your care **(1.1 Formulating learning objectives)**.

Assessment method(s)

Choose an assessment method or combination of methods in line with the learning objectives and possibilities (go to **1.2 Assessment methods** for suggestions).

Step 2) Designing assessments

Assessment scheme and test specification table

Beforehand. Draw up an assessment scheme and an test specification table (for a more detailed explanation and an example, go to 1.3 Assessment scheme and test specification table. **1.3 Assessment scheme and test specification table.**

For the sake of sake of transparency, make the scheme – stating what assessment method(s) will be used, weighting, conditions – available on Canvas.

Designing written tests

Construct the test, while keeping in mind the most important rules for creating written tests (go to **1.4 Designing written tests** and **Appendix 1: Tips for written tests with closed (MC) questions** and **Appendix 2: Tips for written tests with open-ended questions**. Design an answer key (for closed questions) and/or an answer model (for open-ended questions) with a scoring system (points for each [sub]question).

Ask a colleague to check for wording and ambiguities. A colleague can also help with estimating how much time each question will take, to estimate the total time needed for students.

Transparency: put sample questions or a trial or mock test on Canvas, so that students will know what to expect (in terms of the types of questions, their difficulty, etc.). If possible and preferably, discuss the questions and answers during class.

Designing assignments

Draw up the assignment and determine the shape it will take (individual assignment, work in pairs, in groups, etc.). Design the process (handing in draft versions or not; feedback options; deadlines; completion date; the role of supervisors and evaluators, etc.). Establish clear assessment criteria that are in line with the learning objectives. Mention the requirements (e.g. 1500-1800 words, deadline) an inform the students about the consequences if they are not met.

Transparency: Make the assignment, alongside a description of how it will be assessed and the assessment criteria (or rubric), available on Canvas. Check with your students whether the assignment, the process and the requirements are understood.

Step 3) Organizing the examination – written test

Administering a written test

For guidelines with respect to administering written tests, check the Education and Examination Regulations (EER) of your educational programme and the Rules & Regulations of your Examination Board For instance to know how to deal with fraud.

To read more about what to consider when organizing a written test, see: Assessment support toolbox (Organize test taking). This site, provides also information about the options for digital assessment. The UT Examination Office arranges a lot for you, such as scheduling a suitable location and setting up the Chromebooks (laptops). Check the information on their site to see what they can do for you and what is expected of you, as the responsible examiner.

Step 3) Organizing the examination - Assignments

To organize the assignment process

If there are multiple people involved in assessing and grading the assignment, it is important for everyone to be properly prepared for the process (regarding, for instance: when and how feedback will be given; deadlines; any conditions and consequences that may apply, such as for not handing in draft versions on time) and to be aware of the assessment criteria, ensuring everyone will be able to give proper feedback. Develop an assessment protocol (who assesses what; the role and shape of interim feedback; how the results are determined; any conditions and consequences that may apply, such as for not handing in the definitive version on time, etc.).Other focus areas: preventing and detecting fraud (such as plagiarism, unallowed use of AI) and, in case of group assignments, how to detect and deal with free riders.

Step 4) Assessing the results

Assessing the results of written tests

For open-ended questions: score the questions by way of an *answer model.* If there is more than one assessor, it is a good idea for all of the assessors to discuss the answer model amongst themselves beforehand and in case of doubt or borderline cases, have two separate assessors look at the results.

MC testscan be checked automatically if digitally administered. Or at least easily by having an overview of all the 'keys' (correct answer codes).

Cutting score and grades

Prior to awarding a grade based on the scores, a *cutting score* – which score yields to the passing grade 5.5 - is to be decided upon. For tests with MC questions, the probability of guessing the right answers to these questions can be taken into account. In order to determine the grades, you can create a file, for instance in Excel, with each student's results for all questions, and calculate automatically the total scores and the accompanying test grades.

If tests are reviewed automatically (Remindo, Contest) you will receive an overview of the data automatically, and you can also opt to have the grades calculated automatically as well.

For more information, go to: **1.5 Determining cutting score and grades**).

Assessment of assignments

For assessing assignments, determine assessment criteria or create a rubric. Decide and describe how the grade will be determined based on the scores. In an assessment protocol you can establish who will be involved in the assessing process; how the results are determined; what will be done to ensure inter-rater reliability; whether and how a plagiarism check should be done; any conditions and consequences that may apply if conditions are not met, such as for not handing in the definitive version on time, etc.).

Determining grades for assignments
The results on criteria level can be displayed afterwards in various ways, not just showing the overall grade, but also split results for individual students and criteria. This enables highly specific feedback, gives assessors a chance to compare their scores for each individual criterion and is useful for assessment analysis afterward.
Before determining grades, you first need to determine the method of converting scores into final grades and set the cutting score. Certain weighting factors may play a part in converting scores to grades, and certain elements of the assignment or certain criteria may have been deemed to be crucial (i.e. these have to be met or passed in order to get a passing mark). Certain conditions or requirements may apply, for instance deadlines, and consequences can be plied.

Overall assessor reliability
All of the assessors involved are supposed to perform their assessments similarly. In order to be able to do so, clear criteria and coordination (awareness of and the same interpretation of criteria; calibration of scores and marks) are required. For calibration purposes, it is a good idea to have at least 10% of the final assignments assessed by a different assessor, especially borderline cases. Afterward, based on the student results, preferably on criteria level, it can be checked whether there are significant differences in the results and whether this is explainable.

Step 5) Performance analysis and test analysis

Performance analysis

Once the examination results are in, compile an overview of the grade distribution. Determine the number and percentage of passes and the individual grades (frequency, variance, etc.). Reflect on these results. Do the results match up with what you expected? Are there any peculiarities, and are you able to account for them?

Test analysis on item level

Performing a test analysis on item level prior to grading is very important because it will show whether the test was in any way deficient and whether any immediate measures may need to be taken. For instance, it turns out that two answer options were correct for a MC question.
A test analyses is also helpful for evaluative purposes. It provides an indication of the quality of your examination and of the teaching process. This may give you ideas for future improvements.

Test analysis: Perform a psychometric analysis before determining the final mark and take suitable measures if necessary (for a more detailed explanation, go to 1.6 Test analysis).

*Test analysis for tests with open-ended questions: R*eview the scores for each individual question (or calculate their *p*-values – go to **5. Test analysis** for more information). Which components do students not sufficiently master? And what do they excel at? While marking exams, keep track of any mistakes frequently made. Are there any reoccurring mistakes that stand out, or any misconceptions?

Test analysis for assignments: Review the total and/or average scores for each individual criterion. Which criterion stands out? Which learning objectives or skills do many students still have insufficient or poor mastery of? What was done very well?

Student evaluations

The results of student evaluations (written surveys or a panel discussion) can also provide information about the examination, such as how difficult or easy a test was perceived to be, or whether the students felt they were well prepared and informed about the examination (transparency).

Overview of basic principles and focus areas for assessment screenings

Basic principles	Focus areas	Demonstrable through	Criteria
Assess the material/skills your students are meant to have mastered/ acquired by now.	• Clear learning objectives. • All learning objectives are assessed at an appropriate level and in an appropriate manner. • Weighting factors and compensation rules represent the importance of the learning objectives and ensure that an accurate representation is given of the extent to which the learning objectives have been met. • The assignment and assessment criteria are in line with the learning objectives.	– Learning objectives. – Assessment scheme (in case of multiple assessment methods) – Test specification matrix – Test questions / assignments	Validity
The assessment method is suitable.	• The studied material is assessed in a way that fits the learning objectives and aligned with the teaching activities and materials, taking into account any efficiency considerations.	– Assessment methods and justification	Validity
The assessment is of good quality.	The assignment has been formulated clearly. Expectations and requirements have been made explicit. For written tests: The questions - and answer options for MC questions - are of good quality (comply with the rules for constructing tests) and the instructions provided with the test are clear. For written tests: there is a sufficient number of questions to ensure reliability. A limited number of questions have been used in unaltered form from previous exams.	– The test or assignment itself – Peer review beforehand Indicators: • Performance data • Test analysis results • Qualitative analysis of produced work, answers to open-ended questions • Student evaluation data (constructive alignment, transparency, degree of difficulty of assessment)	Validity, reliability
Assessment took place in an adequate and reliable manner.	• For written tests, an unambiguous answer model and scoring system is available. • For assignments: criteria have been determined or a rubric with a scoring system has been developed in line with the learning objectives. • The cutting score has been duly chosen and makes a good distinction for the quality of the students' work, particularly when determining pass/fail. • The conversion of scores into grades is justifiable and correct. • For assignments (this does not always apply): there are multiple assessors for each product (work in groups, individual work). • If multiple assessors, attention is being paid to inter-rater reliability. • No assessing errors have been made. • For assignments: The feedback process has been communicated and grades were justified. The interim feedback was helpful and supported the learning process and final product.	– Answer model / criteria or rubric with scoring system for assignments. – Cutting score incl. justification – Assessment protocol (scoring system; converting scores into grades) + any explanation or justification required. – Description of feedback process, feedback stages, type of feedback. – Description of the coordination process with other assessors regarding the assessment methods. – Completed assessment forms Indicators: • Performance data • Test analysis results • Qualitative analysis of produced work, answers to open-ended questions • Student evaluation data / complaints	Reliability
The students knew what to expect.	• Students have been sufficiently informed regarding the method of assessment and how the grades are determined. • Written tests: Students have been sufficiently informed regarding the types of questions • Assignment: Students are well informed about the criteria and requirements.	– Overview of information that has been provided to the students along with justification by the teacher. – Student evaluation data / complaints Indicators: • Performance data / test analysis data (data may indicate transparency issues)	Transparen-cy
Students are given feedback regarding their results, which contributes to their learning process.	• Students receive their grades and are given the opportunity to inspect their marked exam. • With assignments, students are given feedback and/or a chance to discuss the acquired grade.	– Explanation and justification by the teacher	Feedback/ Stimulating the learning process
Examination was con-ducted properly. There were no peculiarities that may have affected the examination (test taking) or the assessment afterwards.	• Any problems that occur while the test is being held or other peculiarities (such as accidents, calamities, etc.), which may affect the examination or its assessment (for the whole group or individual students). • Instances of fraud. If there are any peculiarities, the related details and the way they are handled are recorded.	– Justification by the teacher (and supplemented by the Examination Board if need be). Indicator: • grades deviate strongly from teacher's expectations or the students' past grades.	Reliability Various

SUPPLEMENT

The following explanatory supplement will provide you with a more detailed explanation of the various topics mentioned before in the Step-by-step guide. This supplement is still being expanded upon. Currently, it offers more detailed explanations for the topics underlined and highlighted in blue (internal hyperlinks in the Step-by-step guide).

Subjects:

1.1 FORMULATING LEARNING OBJECTIVES

What are learning objectives?
Learning objectives describe the knowledge, skills and attitudes that students are meant to have acquired by the end of an educational component. Knowledge in this sense refers to the information they have stored mentally, while skills refer to intellectual as well as physical activities.

Why do we use learning objectives?
Learning objectives form the basis of teaching. When you know what you are trying to achieve, you are better able to determine how to obtain those results (education planning and execution) and how those results may be measured (examination).
Working with learning objectives gives students a clear idea of what will be expected of them by the end of the course.
Learning objectives give direction to teachers' teaching processes as well as the students' learning processes.

How do I formulate learning objectives?
A learning objective must communicate what you expect your students to achieve. Good learning objectives are specific, demonstrable (measurable) and feasible. By incorporating the following elements, you will be able to formulate good learning objectives:

- **Active verbs**: The use of active verbs (along with informing students of the intended result) gives students a clear idea of what is expected of them and the skills they must be able to demonstrate by the end of the educational component. For example: 'By the end of the course, the student can *name the characteristics of …'*
- **Conditions**: The conditions indicate the circumstances or requirements, such as anything that the students can or need to make use of or include (theories, etc.) and/or the situation or context ('… for a simple problem of the following kind…') and/or the material that they will need to apply their acquired skill to (new/known material).
- **Standard:** The standard indicates which criteria, standard or level will apply. This is not always made explicit. For example: 'The student can … *without any measurement errors.'*
- **Results**: the intended result under the specified conditions. NB. A learning objective will lead to a demonstrable result, but the form need not always be made explicit. For example, if the student needs to be able to explain something, the intended result will obviously be an explanation.

Example
Do*: The student can explain the influence of the interface on the user, using cognitive ergonomic concepts.*
The student can explain a given user problem from the perspective of the relationship between man and machine.
Don't*: The student knows about the relationship between man and machine.*

For a nice video about learning objectives, see: https://youtu.be/eXxTpDg1thI

Note on the interrelationship between learning objective, teaching and examination
A properly formulated learning objective indicates *what the student is able to after following your course.*
The first example mentioned above indicates that the student needs to understand (*the student can explain)* cognitive ergonomic concepts. For the second example, the teacher will focus on providing

examples of and practicing the application of knowledge regarding the relationship between man and machine to teach students how to explain user problems.

A properly formulated learning objective gives direct insight into what will be assessed and how it will be assessed. Learning objective: '*The student can explain a given user problem from the perspective of the relationship between man and machine.'* For assessing this learning objective, the students will be likely presented with a user problem (a new problem, but similar to problems used in exercises) and be asked to use the knowledge and insights they acquired to explain the problem. The (measurable) result may be requested orally (oral examination) or in written form (written test).

Active verbs
Using active verbs allows you to properly indicate the intellectual level that students must be able to demonstrate. We can distinguish between knowledge, insight, application and problem solving:

Knowledge:

The student is able to list the characteristics of ....
The student is able to give the definition of … The student is able to select .... from ....

Insight:

The student is able to explain how ….
The student is able to explain …. in their own words
The student is able to describe a situation in which the following applies .... The student is able to explain the differences and similarities between ….

Application:

The student is able to choose the right method for …. The student is able to apply the solution provided
The student is able to design …. using the procedure provided
The student is able to determine ..... by ….

Problem solving:

The student is able to outline the route to a solution The student is able to analyse .... for ....
The student is able to find the error in....

Taxonomies as a helpful tool
Taxonomies can be very helpful in creating learning objectives (and questions for a written test). The most well-known taxonomy for learning objectives is developed by Benjamin Bloom, but other varieties exist.
Some interesting sources of information on this topic are the following:

- Creating learning objectives - John Cline (video): https://www.youtube.com/watch?v=_woMKwBxhwU
- Bloom's Taxonomy: Why, How, & Top Examples - video
- Using Bloom's Taxonomy to Write Effective Learning Outcomes | Teaching Innovation and Pedagogical Support Explanation. University of Arkansas

1.2 ASSESSMENT METHOD(S)

The guiding principle for choosing an assessment method should be your learning objectives. The question that needs to be asked is: Which method is most suitable for allowing the students to demonstrate their mastery of the learning objectives?
A course or unit (and a module normally) may require more than one assessment method, such as both a written test to assess the knowledge of the students and an assignment to demonstrate the practical application of what was learned.

It may not always be possible to use the most ideal assessment method, because of certain constraints such as the number of teachers available, or because the method is not practically feasible. In those instances, it is important to choose an efficient method that will do justice to the learning objectives as much as possible. There may also be additional guidelines or agreements in place regarding the methods to be used within a specific programme. These will be stated in the programme assessment policy, EER or R&R of the Examination Board.

Formative and summative assessment
Summative assessment is applied to arrive at a formal grade and/or final decision on whether the student has sufficiently mastered the learning objectives of a given course. The results (usually in the shape of a grade or simply a 'pass') have certain consequences and are often called 'high stake exams'. All of these 'exams' together are used to assess whether the student meets the graduating requirements of an educational programme.
Because the assessments could have serious consequences for the students and because our aim as a university is to ensure that students meet the learning objectives (and ultimately, the programme aims), it is important to implement proper examination practices. That is: Using suitable methods and considering and taking into account the quality criteria for good quality assessment practice: validity, reliability and transparency.

Formative assessment can be applied during the teaching process and is mainly aimed at giving both students and teachers an idea of whether the learning process is proceeding well (i.e. to what extent the students are mastering the learning objectives) and if any adjustments need to be made to foster and guide the learning and/or teaching process. It also helps students better understand what is expected of them and what the real assessment will be like (transparency). Formative assessment usually involves a feedback element, but the feedback can also be given by for instance peers. Self-assessment (e.g. based on a checklist) can also be seen as formative assessment. Examples of formative assessment: question-and-answer sessions about the material during class, a (digital) quiz, a (digital) mock test halfway through the term, including disclosure and/or discussion of the answers afterwards, a practice assignment that is discussed afterwards or further elaborated on, etc.

Overviews of assessment methods:
> Assessment support - See under Choosing a suitable assessment method
> Alternative assessments | Assessment support

1.3 ASSESSMENT SCHEME AND TEST SPECIFICATION TABLE

An assessment scheme

An assessment scheme is an overview of *all of the assessment methods you will use for your course/unit or module.* It indicates which assessment method will focus on which learning objective(s). It also lists the weighting factors that apply and any compensation options or conditions.
NB. There may be an overlap in terms of the learning objectives covered by the assessments. A certain learning objective may require both a written test and an assignment, with the written test designed to assess the students' knowledge, and the assignment intended to assess their application of that knowledge to a given case study.
For the sake of transparency and managing students' expectations, the assessment scheme should be made available to the students on Canvas. The scheme may also be discussed during class, to really ensure that students are aware of how the examination for the course will take place.

Example of an examination scheme

Name of assessment	**Learning objectives** **By the end of this course, the student will be able to ….**
Partial written test A1	LO1: … LO2: … LO 3: …
Partial written test A2	LO 4: … LO 5: ….
Assignment Y	LO 3: … LO 6: …

Name of assessment	**Assessment method**	**Way of assessing results**	**Individual/ group**	**Materials**	**Weight-ing**	**Conditions**
Assignment Y	Assignment – Apply theories to a given case; Product: report	Rubric	Duo	Book: chapters 1,2; handout X	20%	≥ 5.5
Partial test A1	MC test, digital with Contest	Key answers, contest form	Indiv.	Book: chapters 7,9; guest lecture	30%	Compen - sation allowed
Partial test A2	MC + open-ended question test	Paper version - Key answers / answer model	Indiv.	Book: chapters 3,4,5 and 6; video	50%	

A test specification table

A test specification table is usually applied for written tests. It gives an *overview per test* of how the learning objectives are assessed. It mentions the weighting factor and – but not per se, the material to be assessed. It also shows what type of questions are used and provides an indication of the level of the questions (using a taxonomy, for instance Bloom's taxonomy or Miller's pyramid; see 1.1 for Bloom's Taxonomy and the box below the table for other options..
Using a test specification table allows you to be sure on the one hand that all of your learning objectives are being assessed (validity), are assessed at the right level, and on the other hand that every time you create a certain test for a particular course, it will have the same composition and level and be of the same quality. This is most of all important in the event of resits (*Tip: create the initial test and the resit test at the same time).*
If a database is used, things like learning objectives, knowledge level, degree of difficulty, etc. can be coded, making it easier to create an equivalent test every single time.

Step-by-step guide to create a test specification table

1. Based on the assessment scheme, select the learning objectives that go with this particular test.
2. List the relevant learning objectives in the first column.
3. In the second column, specify the type of question to be used for assessing each learning objective.
4. In the third column, specify the level at which each learning objective will be tested (e.g. knowledge, insight, application or problem solving). The Bloom Taxonomy can be used as mean, see 1.1 and some sources mentioned below the table.
5. Determine the weighting (in percentages) of each individual learning objective (how heavily they will feature in the overall test), and fill out the percentages in the final column.
6. Based on the weighting factor, determine the number of questions and the corresponding number of points.
7. * Not per se needed: For each learning objective, specify the material that goes with it. This allows you to 'check' whether the learning objectives and the material are well matched.

Example of a test specification table for a MC/open-ended question test

Learning objectives **By the end of this course, the student will be able to …**	**Question format**	**Level**	**Questions (codes, numbers)**	**Points / score**	**Materials**	**Percen-tage of total score**
1….	MC	Knowledge	1,2,3,4,5	5 (each 1 pnt)		5%
2 …	MC	Knowledge	6,7,8,9,10,11, 12,13,14,15	10 (each 1)	Book: chapters 1,2	10%
3 …	MC	Knowledge	16,17,18,19, 20,21,22,23, 24,25	10 (each 1)	Book: chapters 3,5 Article 1	10%
4 ….	MC	Insight	26,27,28,29, 30,31,32,33, 34,35,36,37, 38,39,40	15 (each 1)	Book: chapters 4,6 Articles 2 and 3	15%
5 …	Open-ended	Analyse (1) Evaluate (2)	1a,1b,1c,2	1a-c: 4,3,3 pnt 2: 10 pnt	Book: chapter 7, handout X	20%
6 …	Open-ended	Insight (3) Apply (4-6)	3a, 3b,4,5,6,7	3a: 4; 3b: 6 4,5,6,7: each 5	Book: chapter 8, lecture slides	30%
			Total number of points	100		100%

Some useful sites about taxonomies to describe the level:

- Educational Taxonomies with examples, example questions and example activities. http://www.homeofbob.com/pedagogy/theories/taxonomies/blomstax.html
- Writing Multiple-Choice Questions for Higher-level Thinking. https://www.learningsolutionsmag.com/articles/804/writing-multiple-choice-questions-for-higher-level-thinking
- A Taxonomy for Learning, Teaching, and Assessing: A 3-dimensional Model: RevisedBloomsHandout
- Miller's pyramid Educational theories you must know. Miller's pyramid. St.Emlyn's
- SOLO Taxonomy Solo Taxonomy by Inspiring Inquiry

1.4 DESIGNING WRITTEN TESTS

Important basic principles for proper examination include the following: the test assesses the knowledge/skills that the students are meant to have acquired; the test is able to make a clear distinction between 'strong' students and 'weak' students; and the questions are of good quality (i.e. they are clear and unambiguous).
Proper examination hinges on clearly formulated learning objectives (for more information, go to 1.1 Formulating learning objectives).
For written tests, the test specification table is your blueprint for designing the entire test; in it, you specify how the individual learning objectives will be tested. It offers you a guarantee that all of your learning objectives are being assessed, and on the right level too (knowledge, insight, application). For assignments, this safeguard is provided by the assessment criteria or by a rubric.
If you want to construct an item bank for written (MC) exams, it is important to produce high-quality items and to review it afterwards by means of an item test analysis.
For tips on creating good open-ended or closed test questions, check out Appendix 1 (closed questions) and Appendix 2 (open-ended questions).

Important focus points for designing exams:

- **Unambiguous phrasing of questions, assignments and instructions.** The students must be able to understand what (type of) answer or performance is expected of them. The answer or performance clearly distinguishes between students that do and do not have proper mastery of the material or the skill cq the learning objectives. The following must for instance be guaranteed: questions are not interrelated; the answers cannot simply be guessed; the phrasing causes no confusion or misconceptions (because of double negatives or the like); there are no trick questions.

TIP: *Ask a colleague to review your test or assignment beforehand and to provide feedback. Consult the rules for creating written tests and tips in Appendix 1 (closed questions) and Appendix 2 (open-ended questions).*

- **Check for relevance and consistency.** With each question ask yourself: is this relevant to the learning objectives and is this consistent with the learning objectives and my teaching? For example, how detailed will your questions be? Are they not exceedingly difficult, beyond what your students might expect? (For example, they are asked to demonstrate application of knowledge when the learning objective and what they have been taught focused on insight instead.).

TIP: Important: make sure students know what is expected of them.

- **Length of a written exam.** The test ought to be of a length so it can reasonably be completed within the allotted time by a student who has studied the material well. In order to achieve this, you need to assess how much time it will take to read and comprehend the questions (and any present answers) and answer them. For example, answering open-ended questions usually takes longer than answering MC questions. For an overview of answer time see Appendix 1 – Tips for written tests with closed questions.
  The number of questions should safeguard the test reliability. For MC tests, use the following general rule; questions with 2 options: at least 60-80 questions // questions with 3 options: at least 45-60 questions // questions with 4 options: at least 40 questions.

NB. You do not need to maintain this as a strict rule if the test uses both open-ended questions and MC questions, or if it is for formative purposes (i.e. does not count or counts very little towards the final grade) or for partial tests if these make up the final mark together and are sufficiently reliable overall.

1.5 DETERMINING CUTTING SCORE AND GRADES

Requirements: a basic file (e.g. Excel) with all of the results, grouped both per student and per question or criterion. You can then determine the total score for each student as well as their grade.

Determining cutting scores

Your (written) test will award all of the students a certain number of points through adding up the points they received for each question (score). But how do you convert this score (points) into a grade? What should the cutting score be? What should constitute a fail, and what should constitute a pass?

Absolute and relative cutting score
The cutting score is the boundary between pass and fail when assessing a test or assignment. You determine a minimal score that is required to pass; that is: the score that will yield to grade 5.5.
There are three basic methods: absolute, relative, and a compromise of the two.

With an ***absolute cutting score method,*** you determine the boundary between a pass and a fail beforehand based on the learning objectives. How much do you think the student should know/be able to in order to pass? Which score or percentage (criterion) represents this degree of knowledge or ability? For example, students must answer a minimum of X questions correctly, or obtain at least 55% of the total possible score. NB. Often 50% or 55% is chosen as the cutting score percentage, but if there are no strict rules indicated, you can choose otherwise. For instance, for a driver license the cutting score is much higher for a pass/fail. For the Insight questions you need to have 89% correct.
For MC exams, it is advised to take the chance or guess factor into account, which will increase the cutting score.

With a ***relative cutting score method,*** you use your students' results as a basis. You compare a student's score to those of his fellow students and base your grading on that. For example, you give the top 10% of students a 10, the next 10% a 9, etc.

You can also use ***a compromise*****, such as the Cohen-Schotanus (1996) method**. This method uses an absolute cutting score initially, but then also takes into account the degree of difficulty of the test. For this method you take not the theoretically highest attainable score as reference point, but the actual highest score or the average of the students scoring in the 95th percentile and above. score-digit transformation scale of figure would then include, for example, not 100 as the highest score, but 93 because that was the highest score achieved.

At the UT we generally use the absolute cutting score. If the programme's Education & Examination Regulations or the Rules and Regulations of the Examination board do not prescribe a specific wat to decide about the cutting score, there is freedom in choosing the way you determine the grades, but be aware that you have to be able to justify your choice.

Determining cutting score and grades for MC exams
The absolute cutting method is used as the basic principle in this instance.
When determining the cutting score for MC exams, **you can take into account the guess factor.** You take into account the probability that students can guess the right answer without having the expected knowledge, insights and skills.
How do you convert scores into grades for MC exams?

Example:
Consider the following: an MC text with 40 questions (=n), with 1 possible point for each.

Determining the cutting score
In that case, the cutting score (threshold between a pass and a fail) will be determined as follows:
Cutting score = nr + [(n - nr) x p]
nr = number of questions (score) based on guessing (40/4 = 10) n = highest possible score (40 points)
p = the required knowledge proportion or required % of answers that must be correct for a pass (55% or 0.55).

This puts the cutting score at: 10 + [(40 – 10) x 0.55] = 26.5
Scores of 27 and up will result in a pass for the written portion of the exam. Scores of 26 and under will result in a fail for the written portion of the exam.

Example of using the Cohen-Schotanus cutting method for closed questions
Consider the following: the quality of the test questions was good, but in hindsight it appears that the test was very difficult. The student with the highest score answered 35 questions correctly. Let us refer to this value as n' *(NB. Instead of using the highest score as a benchmark, you can also use the average score of the top 5% of the students, i.e. those in the 95th percentile and up.).*

The new formula for determining the cutting point using the Cohen-Schotanus method is then:
nr + [(n' - nr) x p].
For the above example, the cutting point would be: 10 + [(35-10) X 0.55] = 23.75

Using several different types of MC questions
If you are using different types of multiple-choice questions, you add up the probability of guessing the right answer for each of them and use that aggregate in the formula.
Consider the following: your test is comprised of 120 questions (n) with 1 possible point for each, 80 correct/incorrect questions, 30 questions with 3 options and 10 questions with 4 options.
The guessing probability of the written portion of this test would amount to:
(80/2 + 30/3 + 10/4) = 40 + 10 + 2.5 = 52.5 (nr).

Another alternative
In case of an equal distribution of various types of questions, it may be fair to say that the questions that are easier to get correct by guessing (2 options) and those that are more difficult to guess at (3, 4 or more options or matching variations) cancel each other out, and that a guessing probability of 25% would be a fair average.

What to do in case of a combination of MC and open-ended questions
One way is to calculate the scores and grades for the MC questions (taking the chance factor into account) and the open questions separately. Keep two decimal places until you count both grades together and take the average (or you may apply a weighting factor).
Another way is to use the guessing factor as calculated for the overall score.

Interesting read about the combination of open and closed questions: Combining closed and open questions in a test – Draaijer on Assessment and Testing

Determining grades

Transformation from score to a grade
If you start with deciding on the cutting score, you need a formula for the transformation from the scores to the grade. On this website you will find more information how you can make this transformation: Grading written tests | Assessment support

Instead of starting with a decision about the cutting score and make the transformation based on this, you can also use a formula to calculate the grades. One method to come from a score to a grade which is also in secondary education in the Netherlands quite usual, is to use the following formula:

grade = (p/t * a)+ b
p = points achieved by a student (or his/her score)
t = total points available
a = grade-digits to be distributed
b = lowest grade-digit possible
NB. In this situation we don't take the guessing factor In for MC questions into account.

If we take grade **1** as the lowest grade (b), then we have 9 grade-digits left for distribution (a). This will lead to an actual cutting percentage of 50%, so 50% of the total score will yield to grade 5.5.
If we take **0** as the lowest grade, we have 10 grade-digits left and the cutting percentage will be 55%.
The only thing then is that we have to round of very low scores, beneath 0.5 to 1, because for Osiris the lowest possible grade is “1”.
For more explanation and examples, see Grading written tests | Assessment support

Tip: Excel can be very useful to calculate the grades for all scores, but automated programmes for converting scores into marks between 1 and 10 exist as well, for instance:

http://omzettingstabel.faistos.nl/ and at http://cijfersberekenen.nl/

When you use a digital test program, such as Remindo, you can get the grades automatically. But make sure that the way Remindo calculates the grades is the way you want it. You can also export the scores and use your own formula for the transformation. But in all cases: Be able to justify your choice.

1.6 TEST ANALYSIS

Test analysis for assignments

If you use a rubric with criteria and standards with scores, you can calculate the P-value for each criterion criteria the way it is described for open questions.
Else you can look at the total scores for each individual criterion or averages. Which criteria got low scores? Which learning objectives or skills do many students still have insufficient or poor mastery of? Which criterion shows good mastery by most students?

Test analysis for written tests

Required: A file (Excel) with all of the results, grouped both per student and per question. You can then determine the total score for each student as well as their grade, and the total score for each question. The scores per individual question allow you to calculate the p-*value (degree of difficulty). If you like, you can group together certain questions that pertain to a specific learning objective and use the data to determine the scores for each individual learning objective.*
If you are using Remindo or Contest, the data are made available to you automatically. Based on the data, you can then create an overview of the grade distribution for subsequent performance analysis (number of passes; frequency of each grade; variance).

Measures to ensure quality beforehand and afterward

The test results for summative tests determine which students have sufficiently mastered the learning objectives, and which have not. Because the results thus have considerable consequences for the students, it is important that the test valid (did the test really test what it was supposed to test, and at the right level?) and that the results are assessed in a reliable way (are the scores/grades meaningful and fair?).
You can take certain precautions *beforehand* to ensure the test's quality. For example, you can use the test specification table to make sure that all learning objectives are being assessed, and are being assessed in a suitable manner. Taking into account the rules for creating tests and having colleagues review your tests beforehand are also good ways to prevent errors.
By performing a test analysis, you are also able to add a quality check *after the fact.*

Test analysis - corrective purpose

Performing an analysis before grading the test may prevent errors in determining the scores and grades, and as such may contribute to ensuring that the test results are as fair as possible.
A test analysis can give an indication as to whether the test was fair and which specific components (questions) were fair.
For example, you may find that in hindsight, there were two correct answers to a certain MC question. In a situation like this, you can take corrective measures before grading (accepting both answers).
The same goes for open-ended questions; the answers given by students may provide an indication that the question was not clear. You may adapt your scoring or even decide that this question should be omitted for the scoring. NB. When you take measures, due to problems identified afterwards, make sure that students are not disadvantaged by it.

Test analysis - evaluation purpose

Test analyses are not only useful because they enable you to take immediate measures when needed; they also allow you to review the quality of your test and your teaching and use the insights as a basis for future improvement.
For example, if questions about a certain course component (relating to a particular learning objective) were answered noticeably poorly, this may indicate that the students did not grasp that component properly. You can then spend more time teaching that specific component the next time around.

What are some things to consider when compiling a test analysis?

A. Performance data. The number of passes, the distribution of the marks. You look for anything that stands out and if those things can be accounted for. Does the number of passes match up to your expectations? Is the passing rate much higher or lower than 70%? What about the grade distribution? Is it a so-called normal distribution, or does it deviate significantly?
B. For open-ended questions: review questions with exceedingly high or low average scores, and any peculiar or highly common mistakes or misconceptions that may have caught your eye while marking the exams.
C. For MC exams: psychometric data of the entire test as a whole and for each individual question.

Psychometric test analysis on item level

By psychometrically analysing the students' test scores on item level, you can calculate the degree of difficulty and the distinguishing ability of each individual test question, and the reliability of the test as a whole. These quality indicators can be expressed in values, which can then be linked up to set standards. Using this method, the results of a test analysis will give you an indication of the quality of the test questions. If the quality does not meet the necessary standard, the teacher has to reassess the contents of the question. If the content proves to be deficient, the assessor may decide (based on their analysis) to remove the question from the test, adjust the scores, or change the answer in the answer model.

NB: Psychometric standards are very useful for providing indications, but should not be seen as absolute. Nor should the results of psychometric research ever be used as the sole basis for making qualitative statements regarding the test. Look upon it as a signal system. The next step is always to analyse and check what has really happened and whether measures should be taken.

Below, you will find an explanation of how to calculate the reliability of the test as a whole (a) and the degree of difficulty (b) and the discriminating ability (c) of each individual test question, followed by a list of '**corrective measures'** that may be taken.

Take note: If a test consists of two separate components, such as one section with open-ended questions and one with closed questions, a separate analysis needs to be performed for each individual component.

a. Reliability

Test reliability refers to the extent to which the test may be considered a reliable measuring tool, regardless of its contents. Is the measurement meaningful, or might we just as well use a coin toss to assess the students' performance?

The quality of the questions and the length of the MC test play an important role in a test's reliability. Generally speaking, the more questions a test contains, the more reliable it is[1]. A MC test with about 60 questions is usually reliable. That is, if the questions are all of good quality of course. In an ideal world, you would want a group to take the same test twice, to assess its reliability properly. This is not possible. However, **Cronbach's α (alpha)** allows us to assess the reliability of a test after only one examination round. Cronbach's alpha is a way of assessing whether several items should be allowed to form a single scale, using the correlation between the various items.

[1] Of course test reliability also depends on the quality of the questions.

Standards for test reliability
The value for alpha ranges from 0 (unreliable) to 1 (maximum reliability). If a test is judged to be unreliable, the results of that test are basically meaningless. If a test is 100% reliable, it means that the test results at least mean something. Whether they have any truly 'meaningful' meaning is independent of their reliability. True meaning must be found by studying the validity of the test's content (via a test specification matrix, among other things). However, reliability is a prerequisite for content validity; if the test is unreliable, it is also automatically invalid in terms of its content.

The reliability norms are as follows:

0.90 and up	good/very good
0.80 - 0.90	satisfactory/good
0.70 - 0.80	middling/satisfactory
0.70 - under	poor/middling

Generally speaking, reliability rates of 0.60 and under are unacceptable.
As with all psychometric data, these norms are for ***indication purposes only***. To ensure proper interpretation, even when the reliability rate is high, it is always a good idea to also review the psychometric data of all of the individual items and take into account any possible contributing factors. For example, the value of Cronbach's alpha may be negatively affected by homo/heterogeneous of items and group and number of participants and items. It won't be reliable to use for small or very heterogeneous groups or a test with just a few questions or content that is not expected to be coherent (like for instance two subject-contents combined in 1 test).
So yes, it can be used to get an impression of the test in total. But be careful with the interpretation of the Cronbach's alpha.[2]

b. Difficulty of questions

Open-ended questions
The difficulty of an open-ended test question is determined by the average score obtained for that question in relation to the maximum possible score. For open-ended questions, 'right' and 'wrong' are usually not the only options; there is a wide range of variation between the two. For example, if the maximum possible number of points for an open-ended test question is 5, the students' scores may range from 0 through to 5. You can then calculate the matching difficulty level by taking the average score (add up all of the students' scores and divide that number by the number of students that took the exam) and dividing it by the maximum possible score. If the average score for a question is 2.7 and the maximum possible score is 5, the difficulty level of that question is: 2.7/5 =
0.54. This value is referred to as the ***p*-value** (between 0 and 1). A low *p*-value indicates a very difficult question. A high *p*-value indicates a very easy question.
All of the questions on the test must contribute to the summative function of the exam. For an optimal contribution, test questions should aim for a difficulty level of about 0.5.

Close-ended questions (MC)
The difficulty of a close-ended test question (with a right and a wrong answer) is determined by the percentage of students that answered it correctly. This percentage is also referred to as the ***p*-value.** For example, if 60 out of 80 students answer a question correctly, the question has a *p*-value of 60/80 = 0.75.

[2] Source: van Berkel, H. and Bax, A. (3rd revised edition 2014). *Toetsen in het hoger onderwijs* (i.e. 'Examination in higher dukaton'). Houten, the Netherlands: Bohn Stafleu van Loghum.)

It goes without saying that part of the *p*-value of a close-ended test question is determined by the fact that students may be able to answer the question correctly simply by guessing. For example, the probability of getting an MC question with 4 possible answers right is 0.25.

Because of this, the optimal *p*-value for close-ended test questions is the median of the maximum *p*-value (1) and the probability of guessing the right answer. For an MC question with 4 possible answers, the ideal *p*-value is 0.63 ((1.00 + 0.25) / 2 = 0.63).

Tests will always include some variance

One of the aims of a test is to distinguish between 'poor' and 'very poor' performing students and between 'good' and 'very good' performing students. Generally speaking, tests include a wide range of *p*-values. Some indicating difficult questions and some the easier ones, but preferably with a clear concentration of questions close to the optimal *p*-value.

Overview of p-*value norms in (summative) assessment*

Type of question		**Median *p*-value**	**Lower threshold**	**Upper threshold**
Open-ended		0,50	0,25	0,90
Close-ended	2 options	0,75	0,61	0,90
	3 options	0,67	0,50	0,90
	4 options	0,63	0,44	0,90

c. Discrimination power

Each test question must make the best possible distinction between students with a high and low final score (high-scoring and low-scoring students). This is what is known as a question's discrimination ability. A necessary condition for being able to distinguish between students is that not all of them answer the question correctly or incorrectly.

The discrimination ability is determined by relating the scores for a particular question to the overall final test scores. Test questions that are often answered correctly by high-scoring students and less so by low-scoring students are apparently questions that are able to distinguish between these two groups of students. Questions that are answered correctly by roughly the same number of high-scoring and low-scoring students do not have this discriminating ability. Questions that are often answered correctly by low-scoring students and less so by high-scoring students do distinguish between the two groups, but in the wrong way. This last scenario always requires the teachers' attention, because it usually means something is wrong; for example, the answer key may contain an error.

The correlation between the individual question score (item) and the final score (total test score) can also be determined. This is known as the **item-test correlation (Rit)**. The final score is the sum of all of the individual examinees' question scores. If an test consists of fewer than 25 questions, the correlation rate is actually too high, because the question used to calculate the item-test correlation is also part of the final score. For tests with fewer than 25 questions, the final score has to be corrected by subtracting the score of the question you want to use in order to calculate the item-test correlation. This is known as **item-rest correlation (Rir).**

Discrimination index and norms

The item-test correlation has a maximum value of +1 and a minimum value of -1. A value of +1 indicates that all students who scored well on the test answered that particular question correctly. A value of -1 indicates that all of the students who scored poorly on the test answered that particular question correctly. However, these extreme values will occur very rarely in real-life situations.

The stability of the item-test correlation depends on the number of examinees. If there were few examinees (<50), the meaning of the item-test correlation must be interpreted very cautiously. However,

the following general rule usually applies: the higher the discriminating ability of an test question is, the higher its psychometric quality.

Overview of discriminating index norms (item-test correlation)

Values	Classification
0.35 and up	Good/very good Satisfactory/good
0.25 - 0.35	Middling/satisfactory
0.15 - 0,25	Poor/middling
0.15 and under	

A question that is able to distinguish well has a positive item-test correlation, meaning that it was answered correctly by more high-scoring students than low-scoring students.

Interpreting the results of a psychometric analysis and corrective measures

Even if a test has been carefully constructed, it may still be found to be of insufficient quality in some areas after the examination has taken place. If several quality indicators are not up to par, there are ways to remedy the situation, if only slightly.

Below, you will find a list of possible results for the quality indicators described above (a., b. and c.). Each case includes an assessment of whether the examiner in charge ought to take measures in that situation, and if so, what measures it should be.

Note: In deciding whether to take corrective measures for specific test questions, it is always good to take the number of examinees into account. If very few students took the test, the quality indicators may come out poor purely by coincidence.

For a MC test: P*-value is lower than or the same as the guessing probability*

When the *p*-value is lower than or the same as the probability of guessing the right answer, often the answer key for that question is wrong. One of the options intended to fool students (a so- called distractor) has instead been labelled as the correct answer. Adjust the answer key and redo the psychometric analysis.

The question may also be a so-called trick question, causing students to be misled into choosing the wrong answer en masse. These types of questions shouldn't be used and better to remove this question, after all, the aim of a test should not be to set students up. Then analysis must be performed anew and scores and grades should be again calculated.

P*-value is (higher than the guessing probability, but) considerably lower than it should have been*

A low *p*-value indicates a complicated question. Generally speaking, some difficult questions can be part of a test to give students the chance to excel. As long as the questions don't exceed what was taught. Whether the better performing students in general did well will be demonstrated by the question's item-test correlation (Rit). If the item-test correlation lies well within the positive range, there is no issue. However, teachers should refrain from working too many of these difficult questions into a test. A test should not be do-able for only the 'best' students in the group and too many difficult questions will usually have a high fail rate. Will this be justifiable?

If the item-test correlation is very negative (with a low P-value), this indicates that low-scoring students were the ones to answer the question correctly, instead of high-scoring students. In that case, the content of the question ought to be checked. What happened here? Maybe two answer options were correct and the 'high achievers' choose the other one, not seen as correct? Or the question was very confusing. There may be a reason te omit the question.

P-value is almost 1.

A maximum *p*-value of (almost) 1 indicates that nearly every student answered the question correctly, i.e. that it is an extremely easy question. Such a question is not useful in distinguishing between good and poor performing students. It may be also be that something about this question made it easy to guess the right answer even without possessing the intended required knowledge. However, it may also be the case that the subject was understood well by most of the students, which is positive. This score warrants a closer look, but is not a reason in itself to justify removing the question from the test.

Item-test correlation is negative or 0.

A negative item-test correlation indicates that more high-scoring students than low-scoring students answered it incorrectly. This is peculiar. The cause may be an incorrect answer key (option A is set as the right answer, when it should have been option B). Such an error must be rectified right away, and the test analysis must then be performed anew.

Even if the key turns out to be correct, a negative item-test correlation is a signal to check the question thoroughly.

NB. When (almost) everyone has the question correct or incorrect, the Rit will be low, but in that case it is evident, there is no discrimination power in place.

Item-test correlation is positive, but lower than 0.15.

A question with a low item-test correlation does not discriminate well between high-scoring and low-scoring students. In this respect, questions of the sort are not very good, but that alone is not enough to warrant removing them from the test. If the *p*-value of the question is also cause for concern (for instance when it is almost the same as the guessing probability), those two irregularities together are enough to warrant a good look into what happened and maybe removing the question from the test, for more reliability.

Cronbach's α (alpha) is lower than 0.70.

A reliability rate of less than 0.70 is too low to use as a basis for decision making, as it will result in too many incorrect passes or fails.

If a test has such a low reliability rate, a relatively large number of questions will have shown a negative item-test correlation. Sometimes by removing some 'bad' questions from the test, the value will differ significantly. Showing that in general the test was okay.

Another option is to run another reliability analysis for subsets of the items used; for example, you could calculate the *α* values of all knowledge-based questions, that of all insight-based questions and that of all application-based questions, respectively (insofar as the test contains questions of these kinds). If the individual *α* values are sufficient, this indicates that there is no real issue.

But be careful with drawing conclusions based on the Cronbach's α, there are some ifs and buts, see **a. Reliability**

Cronbach's α (alpha) is higher than 0.70, but lower than 0.80.

A reliability rate between 0.70 and 0.80 is middling, and therefore worth examining in more detail. If there are more assessments for a reliable impression, an *α* between 0.70 and 0.80 is deemed sufficient. However, check the questions with a negative item-test correlation who will have contributed to a lower value.

Careful with removing a lot of questions.

You can remove questions with flaws from your test afterward, as long as students aren't disadvantaged by this. This might enhance the reliability, but might also lower the test's representativeness and validity. For example, in case most of the questions about a certain topic are removed. As such, teachers must take into account the representativeness of the test when deciding whether or not to remove questions, and carefully balance the two to choose the lesser of two evils.

APPENDIX 1 – TIPS FOR WRITTEN TESTS WITH CLOSED QUESTIONS

A customary closed test question or MC question consists of:

- the stem (question) or a stem (sentence with necessary information) + a question or a case study + one or more questions about the case. A case study may include a piece of text, a formula, a drawing, a video, etc.
- answer options: the key (right answer) + distractors

Most common question types:

- correct/incorrect question (statement question)
- multiple-choice question / one-of-multiple-options question (question + answer options a, b, c, etc.; between 2 and 5 answer options, with one of them being the correct or best answer)
- multiple-choice insertion question (a sentence is provided with one or more words missing from it, usually at the end; the answer options list the possible words to be inserted)
- more-than-one-option question (more than one answer is correct)
- ordering question (the options have to be put in the right order)
- matching question (two sets of answer options have to be matched into the right pairs; the sets of options may not be of equal length)
- matrix question (a collection of data is provided, and the student has to answer which characteristics do or do not apply to it)

Tip: a test specification matrix will help you make sure that questions are in line with the learning objectives (a suitable number of questions per each learning objective based on their importance and weighting factor). It provides a framework to ensure that you will be able to create an equivalent test the next time around.

Number of questions

The number of questions is determined by:

- the number of questions required to ensure reliability;
- the purpose of the exam; if it counts towards the final mark, its reliability is of greater import;
- is there one single final exam, or are there multiple partial tests(that contain a sufficient number of questions overall);
- the available time;
- the make-up of the exam; only MC questions or open-ended questions as well (in which case a smaller number of MC questions would suffice), or a combination of several types of MC questions.

Overview of answer time (in general) and number of questions required to ensure reliability		
Type of MC question	**Answer time ***	**Minimum number of questions required to ensure reliability**
Correct / Incorrect or 2 options	approx. 50 secs	80
3 options	approx. 60 secs	60
4 options	approx. 75 secs	40
Short case study	approx. 120 secs	
Long case study (1/4 page)	approx. 5 mins	

** The answer time also depends on the question's degree of difficulty (for example, studying a schematic takes a lot of time) and the actions required (such as when a calculation has to be completed first).*

Important focus points when creating questions and possible answers

- ✓ Give clear instructions, particularly for more unusual types of questions.
- ✓ Be comprehensive, but make sure that questions and answers are brief, so that little time is lost while reading. The perfect question can be answered without having to read the answer options and does not contain any unnecessary or trivial information.
- ✓ Divide longer questions up into a stem with the information (or case study) and a separate question. Any text that reoccurs in each of the answer options should instead be included in the stem.
- ✓ Make sure that the question and the answers (and the answers in relation to one another) differ as little as possible in terms of language use and jargon. Make sure that students cannot simply guess the correct answer based on the jargon in that answer option.
- ✓ For one-of-multiple-options questions: make sure that only 1 of the answer options is unambiguously correct (for 'choose the correct answer' questions) or unambiguously the best answer (for 'choose the best or most fitting answer' questions).
- ✓ Make sure that the questions are not dependent on one another.
- ✓ Choose logical distractors, based on common or expected errors in reasoning. Do not create nonsensical distractors; instead, leave these out entirely.
- ✓ Do not include trick questions or questions that cause unnecessary confusion.
- ✓ Underline negations and/or put them in bold text (i.e. 'What should you **not** do in case of a fire?').
- ✓ Choose a specific system for ordering the answer options; for example, always list the options in alphabetical order, or in case of numbers, from smallest to largest. Unless, of course, using such a system would make the correct answer easier to guess.
- ✓ If you include any statements, opinions, quotations, conclusions etc., make sure to specify who said it, which theory or source makes that claim, etc.
- ✓ Be wary of descriptions or answer options that may be interpreted in more than one way or may be hard to interpret; for questions regarding measurements, weights, distances, etc., always include the exact measuring unit.
- ✓ Answer options ought to revolve around the same concept or school of thought. If the answer options contain multiple concepts, the question ought to be restructured into separate correct/incorrect questions for each component.

For both questions and answer options, avoid:

- ☐ Vague wording (such as 'maybe', 'almost always', 'roughly', etc.) or absolute wording (such as 'always', 'never', 'definitely').
- ☐ Grammar or spelling errors, unnecessarily difficult terminology or jargon (unless it is well known to the students). Avoid complex sentence structures. Take into account non-native speakers.

For questions, avoid:

- ☐ Multiple questions/problems in the stem.
- ☐ Negative phrasing of the stem (unless there is a good reason for doing so).
- ☐ Modal words like 'may' in the question, e.g. 'Medicine X may be a suitable treatment for….'

For answer options, avoid:

- ☐ Overlap in the answer options; they must always exclude one another. *So not: 'A prime number is [a] larger than 1; [b] divisible by 1; [c] divisible by itself; [d] larger than 1, divisible by 1 and divisible by itself', with only answer [d] marked as the correct answer.*
- ☐ Using 'all of the above'/'none of the above' as answer options.
- ☐ Negations in the answer options which, when paired with the question, would result in double negatives.

- ☐ Answers containing literal text from the book (students would then choose based on recognizability).
- ☐ Answer options that do not line up with the question grammatically, or instances in which only the correct answer lines up with the question grammatically, thus giving away the right or wrong answer.
- ☐ A considerable difference in answer option lengths (particularly if that difference provides clues as to which is the correct answer).
- ☐ Hints in the answer options that suggest that they are correct or incorrect: for example, a term that is in the question only features in the correct answer option.

Be careful with:

Yes-no or correct-incorrect questions, especially a great many of them

If you are using a lot of correct-incorrect-type questions, make sure that for about half of them, the right answer is 'incorrect'. Creating good correct-incorrect questions can be difficult, because our focus when learning material tends to be on what is correct and not on what is incorrect. Because of this, questions of this type may end up feeling very contrived. Moreover, the incorrect option has to really be unambiguously incorrect, which is not always the case. Another disadvantage is that the incorrect options may become lodged in the student's memory if they are not given feedback.

This question format is valid in some cases, but still requires extra care.

Double statement questions

These are questions that follow this format: Statement A… Statement B… Answer options: (a) A is correct, B is incorrect. (b) B is correct, A is incorrect. (c) Neither are correct. (d) Both are correct. These types of questions require very careful parsing and a great deal of mental effort. Each individual statement has to be examined individually, and errors are easily made. If a student knows whether A is correct, but not whether B is, they will be forced to guess, and in doing so, they might answer the question incorrectly, in spite of knowing part of the right answer, and vice versa.

One precondition is that the statements must refer to the same theme or concept. In some instances, double statement questions may be valid. However, it is usually a good idea to decide exactly what you want the students to demonstrate and then choose a different question format accordingly.

APPENDIX 2 – TIPS FOR WRITTEN TESTS WITH OPEN-ENDED QUESTIONS

Open-ended questions ask for a (brief) description or explanation, a list, a calculation, etc.
With open-ended questions, the question itself must be clear, as well as the instructions regarding the type of answer you are looking for. Questions like '*Do you agree with the above statement?*' are problematic, because they will be answered with 'yes' or 'no' with additional argumentation, which does not provide you with any actual information. Questions like '*How do you feel about the above statement?*' ask for personal opinions, which are technically always correct.
In your instructions, include an answer length limit (visually, by including empty lines or limited space, or by stating a maximum word count). This prevents long explanations and helps make the grading process more efficient.
There are ways to indicate the type of answer you are looking for. For example: list examples of …: 1) … 2) … 3) … For lists, be sure to specify the required number of items (Name 3 characteristics of…) and/or provide an indication of how many points will be awarded or subtracted for each correct or incorrect item.

Tip: while creating a test, make sure to write out the full correct answer yourself, or even start by doing so first. Then review the way the question and the answer relate to one another.

Tip: have one of your colleagues check whether the questions are clear. Do they interpret them as you intended? What kind of answer would they give? This is also useful for finding out how long it would take to answer a given question.

Language use – avoid:
- ☐ Misconceptions because of ambiguous language use or because the question can be interpreted in more than one way.
- ☐ Spelling errors, grammatical errors, complex sentence structures, unnecessarily difficult terminology or jargon.
- ☐ Double negatives.
- ☐ Unnecessary negations; try to use positive wording or accentuate important words.

Information:
- ☐ Provide enough information to enable answering the question, but avoid including trivial or irrelevant information; only provide visual context information (an image, a graph, etc.) if it is necessary for answering the question.
- ☐ Specify whether the students must provide an explanation, argumentation, clarification, etc.
- ☐ Separate the question from the contextual information (case study, problem, etc.), also visually.
- ☐ Specify the maximum number of points that may be obtained for each question, so that students can decide for themselves the order they want to answer the questions in.

Relevance and level:
- ☐ Make sure that the questions are in line with the learning objectives, both in terms of content and in terms of level; to be able to answer the question, the student is required to make use of the material they were supposed to study.
- ☐ Do not use trick questions.
- ☐ Neither the question nor the information provided with that question or with previous questions in the same test contains any accidental hints that may help students answer it correctly.

Presentation:

- ☐ If a question consists of multiple sub-questions (for a case study, for example), ask those sub-questions separately and make sure to clearly distinguish between them (visually, through numbering, etc.).
- ☐ If a question refers to a drawing, piece of text, graph, etc., make sure that the reference is unambiguous, and take into account potential color blindness.

Assessment:

- ☐ To ensure objective scoring and grading, draw up an answer model with a scoring system, and decide in advance how you want to award points in case of partially correct answers.
- ☐ If an answer is not the answer you had in mind, but technically not an incorrect answer to the question, that answer must be marked as correct. For example: *What is a prime number? Answer: a number that is divisible by itself.* You intended to ask for all of the features of a prime number, but this answer, which lists only one, is technically correct. Next time, it is better to make sure to ask questions in a way that ensures unintended or incomplete answers are not correct, for example: *Name <u>all</u> of the features that make a prime number.*
- ☐ If you base your marking in part on the clarity of the phrasing, the structure of the answer, the level of detail, etc., clearly state this. For example, if the answer must incorporate the use of a certain method, mention this method ("Use the method XXXXX to calculate…. and show the results for each of the four steps… ")
- ☐ Handwriting, concise language use, grammatical or spelling errors, etc. may affect your assessment of an answer. Take these biases into account while marking.
- ☐ Be aware of innate flaws in marking caused by your knowing the test inside out, knowing whose test you are marking, etc. For example, mark tests anonymously. If your method is first marking all first questions, then all second questions, etc., be aware that an incorrect answer will seem a lot more incorrect to you if you've just seen that question answered correctly in the past however many tests, and vice versa. Fatigue may also have a negative effect on your marking.
- ☐ Generally speaking, it is a good idea to mark all answers to a question first, and then move on to the next question, as it will focus your concentration on one specific question. However, be aware that if after marking a few tests you notice that the answer model is wrong and you adjust it, you will have to review that question in the tests you already marked. Make sure your marking is consistent.