Examination Policy and Guidelines 2014-2018

School of Psychology and Artificial Intelligence



Radboud Universiteit Nijmegen

Acknowledgements

© Many have contributed to this policy document from the School of Psychology and Artificial Intelligence in the 2012-2014 period, including the former Psychology Exam Team (now the Quality Assurance Team). Draft versions of this report have been discussed in the staff meetings of the Psychology Bachelor's and Master's programmes and the Artificial Intelligence programmes and have also been used in external reviews during the 2012-2014 period. At the end of 2013, the examination boards provided recommendations for a penultimate version and a single Examination Policy for the entire School, including, where necessary, programme-specific supplements or exemptions. The Examination Policy was approved by the Director of Education on 15 March 2014.

Examination Policy and Guidelines 2014-2018

School of Psychology and Artificial Intelligence

Table of Contents

Foreword

In recent years, we at the School of Psychology and Artificial Intelligence have made significant improvements in our examination and assessment methods. This has been demonstrated in the following ways:

- educators vary the form of the exams by, for example, combining multiple choice questions with open questions and even adding essay assignments now;
- educators have multiple choice exams screened in advance by the Quality Assurance team and afterwards, have item analyses performed by the Institute for Applied Social Sciences (ITS);
- internship and thesis coordinators are in the process of improving the assessment protocols for internships and theses;
- the second assessor assesses the thesis independent of the first assessor (the 'blind four eyes principle');
- the Examination Boards inspect the quality of exam and thesis assessments.

Thus, we have made significant progress, however, more is necessary. Our efforts to date have been focused on improving the examinations of each programme component. The reasons to now review the exams at the curricular level are:

- Accreditation Organisation of the Netherlands and Flanders (NVAO) requires a description of our examination and assessment methods, which are used to demonstrate that our students have achieved the intended final qualification[1];
- the Executive Board requested, in *Plan van aanpak Toetsing en Beoordeling, Radboud Universiteit, October 2013* (Action Plan for Radboud University Examination and Assessment), that all programmes clearly specify their examination policy and exam programme by 1 March 2014;
- programme and study track coordinators lack the required tools to control coherence and structure in assessment within a programme, study track or programme year;
- pressure on students is mounting and thus exams and assessments have greater significance than they did previously. This leads to, among other things, complaints from students regarding the exams.

[1] NVAO (2011). Limited programme assessment, p. 7.

With the document *Examination Policy and Guidelines 2014-2018 - School of Psychology and Artificial Intelligence,* we wish to provide these additional steps to achieve the aforementioned objectives. In the first chapter, we describe the Examination Policy which applies to the six degree programmes[2]. This part is the most relevant for the heads of the programmes and the year and study track coordinators. However, lecturers/examiners must still be familiar with the information contained in this chapter.
In the second chapter, we describe specific guidelines, especially designed for lecturers/examiners and intended to offer support in implementing the Examination Policy.

For questions about the Examination Policy or Guidelines, please contact the Quality Assurance Team: kwaliteitszorg@psych.ru.nl.

[2] Bachelor of Psychology; Bachelor of Artificial Intelligence; Master of Psychology; Master of Artificial Intelligence; Research Master in Behavioural Science; Research Master in Cognitive Neuroscience.

Chapter 1: Examination Policy

1.1 Objectives and Alignment

The Examination Policy has the following *objectives*: (a) to describe the desired examination and assessment methods, through which the degree programmes can ensure that students achieve the intended final qualification; (b) to provide students with an overview of the examination methods and thereby allow them to manage their learning activities; (c) to allow lecturers and examiners to make responsible decisions regarding examination and assessment, and (d) to provide a framework for evaluation and possible adjustment of the exam quality.

The Examination Policy does not stand on its own, but must be *aligned* with:

- the programme plans of the School's six degree programmes;
- the School's Quality Assurance Policy and the RU's Quality Handbook (Version 3, January 2013);
- the Model Rules and Guidelines of the RU Examination Boards, from 14 November 2011;
- Education and Examination Regulations (OER) of the degree programmes.

In addition, the Examination Policy has, as much as possible, been based on educational and didactic research, assessments from the programmes and best practices of the programmes within the RU and other universities.

1.2 Guiding principles

In this section, we outline the four principles widely used with regard to examination and assessment and discuss the five functions of examination and assessment which can be differentiated. We also briefly describe the three most relevant quality criteria for examination.
In Section 1.3, we translate these principles, functions and quality criteria into specific rules of thumb for examination and assessment.

1.2.1 Basic Principles

1. *A sound curriculum consists of didactically consistent courses, which are aligned with the final qualification*

The curriculum/programmes should be designed in such a way that students who are admitted to the programme are able to achieve the final qualification within the nominal

study duration. Biggs introduced the term *alignment*[3] for the relationship between programmes and final qualifications. We use the term 'didactic consistency'[4].

From a curriculum perspective, the programme should first serve as an excellent build-up to the final qualification (*vertical coherence*) during which study tracks can be differentiated if so desired. Second, the programme should exhibit clear coherence throughout each *programme year*. This *horizontal coherence* should be reflected in the integration of contents and in the education team's shared beliefs that the *academic level* in each year is appropriate and feasible.

There should also be consistency within *courses*: the course objectives are derived from the final qualification, they fit well with the academic level for the year and the study track. All of the lectures, literature, discussions, assignments, feedback, interim exams and final assessment are there to ensure that students achieve the course objectives.

2. Learning objectives, examination and assessment guide the learning process

Research demonstrates that students' learning activities are largely governed by examination and not just by the educational curriculum[5]. Students will make an assessment about the appropriate exam results ('Do we have to know this for the exam?') and adjust their learning activities based on this[6]. If an exam only tests the subject knowledge and learning objectives superficially (e.g. the reproduction of knowledge rather than students making connections themselves; or testing material addressed in lectures rather than skills practiced in project group sessions), then students will only learn superficially, regardless of how often the lecturer attempts to motivate students to study the material more in-depth.

In order to stimulate *deep learning*[7], it is crucial that the exams are about 'what matters most' and 'what they really have to learn'. This means that lecturers must already consider what will be tested prior to the start of the course – not as a final element of the course, but as the foundation.

3. A consistent examination programme has horizontal and vertical coherence

From the *curriculum perspective*, the examination programme (meaning, all of the exams) must first form a good build-up to the final qualification (vertical coherence) for each *study track*. Second, the examination programme must demonstrate coherence

[3] Biggs, J. (1999). Assessing for learning quality: Buckingham: SRHE and Open University Press. Obtained from: http://teaching.polyu.edu.hk/datafiles/R131.pdf

[4] Derived from Huisman, W. (2012). Didactische consistentie: zelfstudiemateriaal voor docenten. Obtained from: http://www.iowo.nl/icto/elem/63/

[5] Ramsden, P. (1992). *Learning to teach in higher education.* London: Routledge; Van der Vleuten, C.P.M. (1997). Beyond intuition. *Tijdschrift voor Hoger Onderwijs, 15*(1), 34-46; Cilliers, F.J., Schuwirth, L.W., Adendorff, H.J., Herman, N., and Van der Vleuten, C.P.M. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education, 15*(5), 695-715.

[6] Elton, L. (1987). *Teaching in higher education: Appraisal and training.* London: Kogan Page.

[7] Biggs, J. (1999). Assessing for learning quality: II. Practice. In: *Teaching for Quality Learning at University* (pp. 165-203). Buckingham: SRHE and Open University Press. Obtained from: http://teaching.polyu.edu.hk/datafiles/R131.pdf.

within each *programme year* (horizontal coherence). The vertical and horizontal coherence from the curriculum perspective are shown in Figure 1. [not enclosed in English version].

From the *course perspective,* the exam must be an excellent reflection of the course learning objectives. In a consistent course, the intermediate learning tasks and assignments accordingly reflect the learning objectives and therefore the exam. Thus, the entire course design is geared to achieving the learning objectives and to passing the exam. This coherence is illustrated in Figure 2 [not enclosed in English version].

4. The examination programme uses the available time and resources efficiently
Developing a consistent examination programme would not exactly be an art if time and money were unlimited. But, of course, that is never the case and hence, the fourth principle: feasibility or efficiency, meaning that the examination programme must be achievable with the time and money available and must utilise available resources efficiently. Herein lies the greatest challenge for the Examination Policy.

1.2.2. Functions

Besides the four principles described above, five functions of examination and assessment can be found in the literature, which serve as important orientation guidelines for examination policies. These functions are:

1. *Improvement or Development*, also called 'formative testing': students receive interim feedback about what is already sufficient and about what still requires improvement. Feedback is a powerful tool to influence students' learning behaviour[8]. Clear and development-oriented *feedback* allows students to improve themselves toward a desired state;
2. *Assessment*, also called 'summative testing': based on predetermined criteria, an assessment of pass or fail, with various gradations between, is given. There is a clear cut-off point between pass and fail, or, in other words, between capability and incompetence;
3. *Selection*: a pass assessment grants access to the next portion of the programme, for example, a more advanced degree programme or profession;
4. *Qualification*: the pass assessment grants a qualification (diploma) on which additional rights (title, registration in a trade register, etc.) are linked;
5. *Feedback in regards to the quality of education*: based on the exam results, the programme or lecturer determines to what extent the training has been sufficient.

1.2.3 Quality Criteria

In the research literature and in quality assurance, three quality criteria are commonly used for examination and assessment.

[8] Hattie, J. & H. Timperley (2007). The power of feedback. *Review of Educational Research*, vol. 77 (1), pp. 81-112.

1. Validity
The most important quality requirement is content validity, meaning validity in regards to the content of the learning objectives. A valid exam measures what it should measure. If this were not the case, then we must question our statements about the performance of the students in question. Validity is a necessary prerequisite, which precedes reliability.

2. Reliability
A reliable exam is an exam which takes consistent measures. This consistency should be high so that an accurate assessment of student performance can be made. An exam is consistent if the result in another situation or at another time or by another assessor would be the same. The quality of an exam is based on reliability, a necessary but not sufficient requirement.

3. Transparency
An exam is transparent for students if they know in advance how they will be assessed and on what assessment criteria this assessment will be based and if students understand how this exam contributes to their own professional practice.

1.3 System of examination and assessment

In this section, we will translate the guiding principles listed above into specific rules of thumb for examination and assessment of the six degree programmes within the School of Psychology and Artificial Intelligence. On the one hand, these rules of thumb offer clear guidance, and on the other, allow sufficient space for optimal interpretation by lecturers and examiners, given each individual situation.
We will formulate six rules of thumb at the curricular level and eleven rules of thumb at the course level. The rules of thumb at the curricular level are intended for *programme coordinators, study track coordinators and programme year coordinators*. The rules of thumb at the course level are intended for lecturers and examiners. However, lecturers and examiners must be well informed about all the rules of thumb.

1.3.1 Six rules of thumb at the curricular level

Rule of thumb 1: The target level for each programme year is clearly defined
The degree programme and thus the examination programme of the different years of the programme should show a clear build-up to the final qualification. This build-up should be clearly defined and should be familiar to lecturers and students. The question remains whether the current global classifications of academic levels per year (for example with Psychology: B1 = introductory; B2 = broad; B3 = in-depth; MA = specialised and research-oriented; Res. MA = specialised and research-oriented) provide lecturers with a sufficient basis for the precise determination of the level at which to place their exams, and whether they give students a sufficient ‘sense of direction’ for the academic

development they are expected to achieve. Within the School, we strive to exposit the intended academic level per academic year[9].

Rule of thumb 2: Appropriate exam mix at the intended level per academic year
Ideally, the programme in B2 has a different combination and/or a different weighting of exam types than in B1, and the Master's different from the Bachelor's, etc. In that way, more essay assignments and semi-authentic tasks, for example, should be used in the later years. In the cases of multiple exam types per course, over the course of the training years, there may be a change in the weighting between exam types (For example: in B1, the knowledge assessment with a multiple choice exam counts for 70% of the final grade and practical assignments count for 30%. In B2, this weighting could shift to 50/50%).

Rule of thumb 3: Integrative examination per programme year
The examination programme should also be coherent horizontally, meaning *per programme year*. The rule of thumb says that, per training year, there should be some form of integrative examination, in which students' knowledge and skills from different subjects/study tracks are integrated. For example, within the BA Psychology, the horizontal coherence is apparent in the integrative function of the core themes and of OP1, OP2 and OP3.

Rule of thumb 4: Realistic exam scheduling
A further precondition is that the scheduling of the exam is done in such a way that students can be well prepared for the exam. This means that the workload for students is evenly distributed over the degree programme and that the exams do not overlap. In practice, this could mean that some semester subjects will not have a final exam, but consist of summative assessment tasks which must be completed during the course.

Rule of thumb 5: Striking a cost-benefit balance per programme year
Certain types of exams are more time intensive and thus, more expensive than others. When choosing the exam date and the exam type, the costs should be weighed against the substantial benefits: consider the required investment (time, money) against the information obtained through the exam. What knowledge, skills and thought processes are most valuable to us and at what point in the programme? To monitor the effective use of resources, it is important to:

- precisely determine how much time various exam dates and exam types require. For example, a common argument is that multiple choice questions require less time to correct. However, creating, maintaining and renewing valid and reliable multiple choice questions requires a considerable investment of time, due to the minimal number of questions required and because of the desired psychometric quality;
- determine whether there are clever and creative opportunities to 'reallocate' teaching staff time from one task to another, for example, through the use of peer feedback and ICT;

[9]There are several models available, such as the taxonomy by Bloom (1956; Anderson & Krathwohl, 2001), the 'pyramid' by Miller (1999) and the SOLO taxonomy by Biggs (1982; 2007).

- invest the most time and money in the exam dates that are most crucial for achieving and assessing the BA and MA final qualifications;
- allocate time to the improvement as well as the assessment function of exams, for example, by devoting more time to feedback *during* the course rather than evaluation *after* the course.

Rule of thumb 6: Sufficient spread and build-up in formative assessment
The sixth rule of thumb relates specifically to formative assessment and thus, the development function of examination (see 1.2.2). Ideally, each student should receive, on at least one occasion per course, *interim development-oriented feedback.* Such formative assessment thus constitutes a reliable reflection of the summative exams. However, since giving feedback is time consuming, this may not be possible for all programme components. In these situations, it is important that, in a programme year, the feedback moments are purposively selected and *spread* over the whole year.

It is also important that the formative assessment in the course of the programme has an effective build-up. For example, the feedback from the lecturer is gradually supplemented (and possibly partly replaced) by peer feedback from other students and by self-reflection from the students themselves. In this way, students increasingly adopt a broad academic thought process by internalising the 'voices of role models', as it were, and by gradually learning to assess their own performance and development. The ability to receive, process and give (peer) feedback and the development of self-reflective capabilities are important objectives in a number of courses.

1.3.2 Eleven rules of thumb at the course level

Rule of thumb 7: Validity of exam monitoring with an exam matrix
The rules of thumb at the curricular level are mainly meant to improve the validity and, in that sense, are a necessary requirement for the validity at the course level.
At the course level, the exams must be representative of the cognitive learning objectives (content validity) and cover the metacognitive learning objectives of the course. Content validity involves a representative sampling of the material, which goes beyond concepts such as, 'the exams must cover all the learning material', or 'one question per chapter of the textbook'.
The learning objectives, in turn, must fit with the *academic level* for the year and for one or more *study tracks*, depending on the differentiation of the degree programme. In order to provide insight into the assessment validity, an *exam matrix* (or alternatives) is required starting from the 2014-2015 Academic Year (see 2.2).

Rule of thumb 8: Several exam dates and exam types per course
Assessment reliability increases when there are multiple and varied measurements within a course[10]. As a rule of thumb, at the School, we organise it in such a way that each course has *at least two exam dates or exam types* from B2 onwards. This allows lecturers

[10] Milius, J. (2007). *Schriftelijk tentamineren. Een draaiboek voor docenten in het hoger onderwijs.* Utrecht: Utrecht University(Ivlos).

to be better able to assess the variety of learning objectives, and offers students the opportunity to demonstrate their skills using varied measurements.
In the study guide and online study handbook, the relative weight of each exam component in regards to the final grade must be clearly described. This weighting should be in logical proportion to the priorities set in the learning objectives and the scheduled time investment from students on the particular learning objective.

Rule of thumb 9: Exams are substantially 'refreshed' annually
Reusing old exams in their identical form is entirely out of the question, given their rapid circulation via social media. This is particularly relevant in multiple choice and open question exams. Exams must therefore be substantially refreshed every year. That can be done by developing a sufficiently large question pool, changing answer alternatives, creating brother/sister questions, scrambling question and answer alternatives, using case studies differently, etc.

Rule of thumb 10: Assessment types and weighting will be announced to students in advance
The transparency criterion requires that students do not face any surprises. At the School, when planning their studies, students can check (via the study guide) the assessment types and weighting of the examination components to be used in a particular course. This means that the lecturer responsible must ensure that these details are clearly disclosed in time to be included in the study guide (by 1 May for the first semester and by 1 Dec for the second semester), and that they are not changed thereafter.

In the case of papers, internships and theses, the assessment criteria must also be clearly disclosed at the beginning of the course. This is done via the study handbook either online or in printed form just before or at the start of the course.

Rule of thumb 11: Students are assessed individually
Students are assessed individually. Even if courses, internships or theses require students to work together in a group, there are still ways of assessing students on an individual basis. For example, via lecturers' own observations, students' logbooks, oral examinations, presentations and oral or written reflections by individual students.

Rule of thumb 12: Assessors use a transparent assessment model (using open questions and papers)
There are several good reasons to work with assessment models. First, it allows students to adjust their own learning activities and provide feedback for themselves or for *peers.* Second, it increases reliability and minimises differences in the evaluations between assessors. And third, it enables the psychometric quality of the exam to be monitored on this basis.
The exact format of the assessment model depends on the type of exam. In the case of open question exams, a correction model is necessary (see 2.3.2); papers, logbooks,

reflective reports, internships and theses require a more extensive assessment scheme (or *rubric*[11]*)* (see 2.3.3).

Rule of thumb 13: Assessors apply the methods of 'absolute assessment' or of 'absolute assessment with a relative component' (for multiple choice questions)
The 'method of relative assessment' (more or less fixed percentages of the students that 'must pass' or 'must be dropped', regardless of the performance of the students) is not considered desirable in the School. Grades are preferably calculated using the method of 'absolute assessment'. If this is not possible, then the method of 'absolute assessment with a relative component' is applied[12]. Both methods are further elaborated in Section 2.3.1 and in Appendix II.

Rule of thumb 14: Multiple choice exams are adjusted for guessing and the psychometric quality of the exam is analysed.
Multiple choice exams are adjusted for guessing. Students are informed of this in the exam instructions. The implication of this is that students are better served by guessing an exam question they don't know the answer to rather than leaving it blank, which would mean they would be doubly 'punished'.
Psychometric analysis of multiple choice exams is mandatory. The Institute for Applied Social Sciences (ITS) provides lecturers with an exam report which can be used to decide how to deal with multiple choice questions of unsatisfactory quality (also see 2.4.3).

Rule of thumb 15: Exams and assessments are viewed by a colleague (peer review)
Examination reliability is improved by submitting the exam and the exam matrix to a colleague (the 'four eyes principle'). This colleague could be a lecturer involved in the course, or an expert in the field of examination, such as a member of the Quality Assurance team. They provide a 'fresh view' to ascertain whether all the learning objectives at the appropriate level have been addressed and to highlight any possible textual ambiguities (answer alternatives that are too similar to each other, etc.).
In the event that multiple lecturers submit exam questions, it is important that the lecturer with the final responsibility decides on the final product and monitors the validity, reliability and general level of difficulty.
The assessments are also subject to *peer review*, particularly when the pass and fail rates are conspicuously different from previous years and/or if the decision must be made to switch from absolute assessment to absolute assessment with a relative component, or if, due to divergent scores in the psychometric analysis, decisions must be made to remove questions, approve multiple alternatives or other criteria revisions. In such cases, the lecturer responsible consults with an independent colleague, or with the study track coordinator.

[11] Stevens, D.E. & Levi, A.J. (2013). Introduction to rubrics. An assessment tool to save grading time, convey effective feedback, and promote student learning (2nd edition). Sterling, Virginia: Stylus Publishing. For free online resources, see for example: iRubric (http://www.rcampus.com/indexrubric.cfm). Blackboard also has rubrics; see the Blackboard Manual.

[12] Gruijter, D.N.M. de (2008). *Toetsing en Toetsanalyse*. Leiden: Leiden University.

Rule of thumb 16: Internship assessment by an internal assessor; thesis assessment by two assessors independently of each other

Internship and thesis manuals are made available before the start of the internship and thesis. The assessment criteria are specified in these manuals – also in relation to the final qualifications – and their weighting and grading factor is explained clearly. It must be clearly stated whether an internship or thesis will be assessed integrally or separately. If both are to be assessed integrally, then it must be known in advance whether the grade for the internship can be compensated with the grade for the thesis.

An interim evaluation/assessment for internships and theses is standard. The thesis and internship manuals should clearly explain who conducts this evaluation/assessment and on what criteria it is based. It should also explicitly state what happens if the interim assessment does not result in a passing grade.

The internship assessor is responsible for the final assessment of the degree programme internships. The external supervisor can deliver information to help toward this assessment. The assessor will evaluate the internship based on a standard evaluation form.

Theses are assessed independently by two assessors (blind four eye principle) using a standard evaluation form. It must be clear in advance whether only the product is being assessed or the product and the process. If both are assessed, then the assessment criteria for both must be known in advance, including their relative weighting in calculating the final grade.

Discrepancies between assessors are discussed. If no consensus emerges, a third assessor is brought in to decide. Discrepancies of 1.5 points or more and discrepancies between a passing and failing grade are recorded and analysed each year by the coordinator. This analysis may lead to further clarification about the assessment criteria and whether it should be tightened. It may also lead to a fresh collegial discussion between assessors on the shared approach to the assessment criteria.

In order to clearly mark the division between the supervision and assessment of papers and theses, the internship/thesis manual and/or internship/thesis agreement clearly describe how many draft versions students may submit, when the final submission deadline is, and what the possible resit deadline is.

Rule of thumb 17: Examiners evaluate the exams and the assessment in the 'teacher report'

Within six weeks following the final exam, the responsible lecturer/examiner evaluates the process of examination and assessment in the teacher report. It provides general information about the exam, such as the exam forms, the weighting, and the success rate of the first exam sitting. It also outlines what the strengths and weaknesses are in regards to the examination and what possible solutions are available for this. The teacher report, including the exam matrix, is discussed by the programme committee (OLC).

Depending on the evaluation in the teacher report, the design of the exam will be adjusted, where necessary. In the case of exams with multiple choice or open questions, it is advisable to record the psychometric data together with the question in a database to be stored for future examination.

1.4 Responsibilities

Guaranteeing the quality of the examination and assessment process requires optimal organisation of the examinations. This calls for transparency in the division of responsibilities. We are striving for an examination culture, in which each person assumes responsibility and we have open lines of communication with each other.
Table 1 shows the group responsible, at each level, for the didactic consistency between the exam and the education, and thus the validity, reliability and transparency of the examination and assessment.

Table 1: Responsibilities regarding examination and assessment at the School of Psychology and Artificial Intelligence

Level	Responsibility	Assurance & Control	Advisory Role
Curriculum level	Director of Education (at KI, BS and CNS delegated to the heads of programmes and to Master's coordinators at Psy Master's programme)	Examination Board	Programme Committee
Study track/ Year level	Study track or year coordinator	Examination Board	Programme Committee
Course level	Responsible lecturer	Examination Board	Programme Committee

The *lecturer with ultimate responsibility* is accountable for the didactic consistency at the course level. This lecturer also acts as an appointed examiner by the Examination Board and, as such, is responsible for the validity, reliability, transparency and feasibility of the examination and assessment. In courses where several lecturers are involved, the lecturer with ultimate responsibility plays not only a coordinating role, but also a directing and decision-making role in the construction of the exams. Where required, a lecturer with ultimate responsibility can receive support from a course coordinator.
The *year coordinator* ensures that the exams in the relevant programme year comply with the rules of thumb at the curricular level. The year coordinator consults with the lecturer with ultimate responsibility each semester about the alignment of education and examination, on the basis of the teacher reports.
At the programme level, the *director of education* is responsible to ensure the examination programmes of the degree programme are described and regularly maintained and that the examination programmes comply with the rules of thumb or justify and explain circumstances in which a decision is made to deviate from the rules of thumb. The director of education is supported in this regard by programme heads/coordinators (from KI, CNS and BS; and by study track coordinators in the Master of Psychology and the Bachelor of Psychology programmes). In addition, the director of education, or programme head/coordinator, together with the Examination Board, provides the right instruction and equipment to lecturers for examination and assessment and provides appropriate assistance via a *quality assurance team* or comparable support.
The *programme committee* advises the director of education/programme heads on all aspects of education quality, including the examination quality. Teacher reports and related exam matrixes are discussed by the programme committee.

The Examination Board plays a crucial role in monitoring and improving the quality of examination and assessment. The Examination Board does this by setting additional guidelines for the implementation of the Examination Policy, based on the Examination Policy itself and additional questions from examiners or students. The Examination Board also guarantees the completion of the final level, by (a) screening the examination programmes of the degree programme based on the six rules of thumb at the curricular level and by (b) screening random samples of examination material and the accompanying assessments based on the eleven rules of thumb at the course level.
This screening of examination programmes and examination materials occurs once per accreditation period (1x every 6 years) in the Bachelor's programme and twice per accreditation period (1x every 3 years) in the Master's programme in such a way that study tracks or degree programmes are alternately selected.
Additionally, the Examination Board screens a random sampling each year of 10% of the assessments of the Bachelor's and Master's theses (and if applicable, of the internships) from the previous academic year, which were given a final grade of 6 to 10.
The Examination Board employs a multi-year plan for the screening activities, which is disclosed to lecturers. This plan should preferably link with the quality assurance and innovation cycle.

Chapter 2: Examination phases and examination guidelines

This chapter goes into more detail regarding the rules of thumb at the course level. This chapter is intended specifically for lecturers with ultimate responsibility, or examiners. We first describe the relevant examination phases. In the second section, clear guidelines for these examination phases are outlined in order to assist examiners in making responsible decisions regarding examination and assessment, in line with the Examination Policy.

2.1 Examination phases

The lecturer with ultimate responsibility/exam developer/examiner goes through five phases for each exam. Most lecturers go through these implicitly. As the examination quality receives more attention and, more parties become involved, it becomes increasingly important to have a shared vision of these phases and of their planning. These phases are broadly elaborated below.
The *first phase* takes place before the course begins. The final product in this phase is a general design of the exam in which it is laid out what learning objectives are to be tested and how they will be tested. This is also called the exam matrix. Also at this phase, decisions regarding the weighting of the various exam components and the pass/fail grade have already been made, so that this information can be communicated to students.
The *second phase* consists of the exam construction. The exam matrix is further developed into exam questions and an assessment model. The quality of these products is monitored by colleagues and adapted into a definitive exam, on the basis of this feedback. To conclude this phase, all the practical matters surrounding the examination (exam

instructions, sufficient number of copies, etc.) are regulated, so that the exam may be administered in the *third phase*.
In the *fourth phase*, a review of the exam takes place. As objectively as possible, points are awarded for the answers given and these points are translated into grades.
The *fifth* and final phase concerns the evaluation of the exam. In this phase, the lecturer analyses the reliability of the exam, and carries out any necessary changes to the scoring or standards. The information obtained from this step and the findings based on the course evaluation may lead to adjustments of the exam design or the learning objectives of the course. In this way, phase five returns full circle to phase one and the exam cycle begins again. See Figure 3. The evaluation of the exam is described in the teacher report.

Figure 3: *Exam cycle*

Phase 1: Exam design
Phase 2: Exam construction
Phase 3: Administering the exam
Phase 4: Exam review
Phase 5: Exam evaluation

Figure 4 includes a schedule for the examination process, based on a study period of 10 weeks (weeks 1-8: education; weeks 9-10: exams and resits). The figure shows how the exam can be an integral part of the education process. Apart from the learning objectives and the exam type, the weighting, the exam matrix and initial assessment criteria are also already established prior to the start of the course.

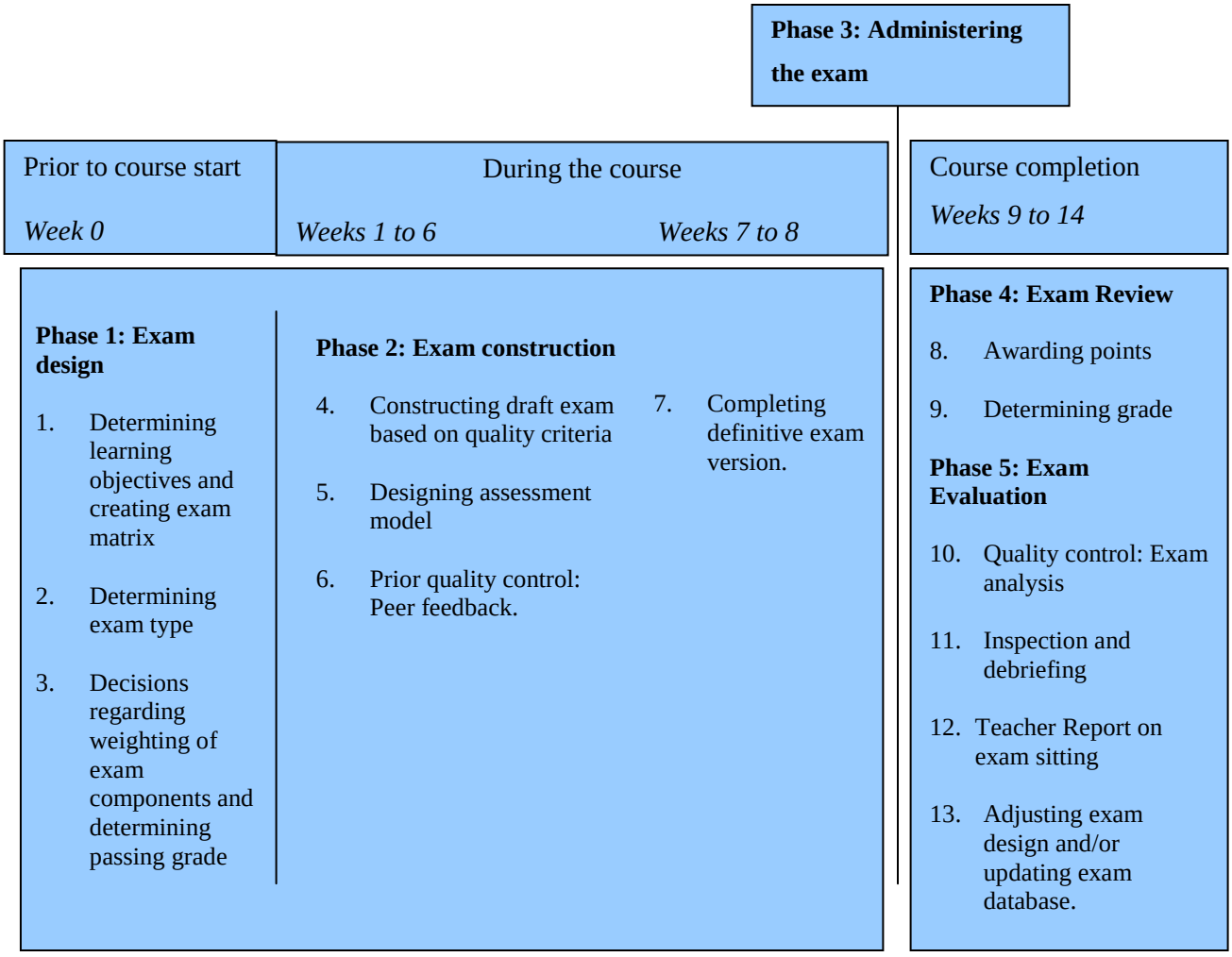


Figure 4: *Five examination phases with corresponding steps, assuming 8 weeks of education and 2 weeks of exams and resits*

2.2 Guidelines regarding course design

In this section, we will look at the guidelines governing the course design (see Figure 4). In 2.3 we will look closely at each exam type in the exam construction and assessment, and in 2.4, we look at the exam evaluation.

1. Determining learning objectives and preparing the exam matrix

To determine the learning objectives, the responsible lecturer looks at the following:

- The final qualifications to which the course contributes
- The learning objectives from previous courses in the study track
- The desired level of the programme year in which the course takes place.

Appendix I may be helpful in formulating the learning objectives.

An *exam matrix* at the *course level* is a two-part schematic overview. First, it makes the relationship between the learning objectives and one or more final qualifications apparent[13].

Table 2: Relation between learning objectives and final qualifications

	Final qualification 1	Final qualification 2	Final qualification 3	Final qualification 4	Final qualification 5	Final qualification 6
Learning objective						
1	X					
2		X				
3	X					
4	X					
5						X

Second, in the *exam matrix* at the *course level*, the relationship between the learning objectives and the exam questions/assignments/cases still to be constructed becomes clear. Two examples are given: a simple exam matrix (Table 3) or a more extensive matrix, in which the cognitive level of the learning objectives is explicitly laid out (Table 4).

Table 3: A simple exam matrix

	Learning objective 1	Learning objective 2	Learning objective 3	Learning objective 4
Question/Task number				
1	x			
2		X		
3	x		x	
4	x			
5		X		
6			x	
7	X			X

[13] This part of the exam matrix is the link to the examination programme (the exam matrix at the programme level), which falls under the responsibility of the programme coordinator.

Table 4: A more extensive exam matrix: the numbers in the matrix represent different questions/tasks; the percentages on the bottom and right represent the share of the final grade; the top row represents the cognitive level

Learning objective	Knowledge & insight	Application	Analysis	Synthesis	Evaluation	Share
LO 1	1,4,17,8	2, 3 etc.				***40 %***
LO 2	5,7,12		10, 16			***20 %***
LO 3	etc.					***10 %***
LO 4						***10 %***
LO 5						***10 %***
LO 6						***10 %***
	30 %	***15 %***	***25 %***	***15%***	***15%***	***100 %***

2. Determining the exam type

The exam type best suited to each course depends on the learning objectives of the course and the chosen teaching format. Common exam types are: exams with multiple choice questions; exams with open questions; exams with papers, presentations and assignments in which students demonstrate skills (such as interviewing and communication skills, statistical calculations, design skills, etc.). We summarise the last exam type under the title *assessments*. The pros and cons of these exam types are shown schematically in Table 4 and identify which exam types lend themselves well to which learning objectives[14] [15] [16].

Multiple choice questions. Exams with multiple choice questions have the major advantage that checking them can be done quickly. A disadvantage of this exam type is that the construction of the exam requires a relatively large amount of time and care. The exam type is therefore particularly suitable for large numbers of students since the time investment for the construction is offset by the time savings in checking.
Multiple choice exam questions are well suited to assessing what knowledge students possess. Multiple choice questions can also be designed so as to test higher *cognitive* skills, however, this requires more attention in the construction process.

[14] Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs.* Houten: Bohn Stafleu van Loghum.
[15] Academisch Centrum Tandheelkunde Amsterdam (1998). *Handleiding tentaminering*. Amsterdam: Academisch Centrum Tandheelkunde. Obtained from http://www.onderwijs.acta.nl/studieweb/docentenwegwijzer/2_tentoe_handleiding_tentaminering.pdf
[16] Vrije Universiteit Amsterdam (2006). *Handleiding toetsen en beoordelen.* Obtained from http://www.fsw.vu.nl/nl/Images/088%20toetsen_beoordelen_2007_tcm30-36518.pdf

Table 5*: Pros and cons of each exam type and link to learning objectives*

Exam type	Pros	Cons	Learning objectives
Multiple choice questions	- Minimal checking time - Efficient with larger student numbers	- Construction is labour intensive	- Cognitive skills at the knowledge, understanding and application level.
Open questions	- Challenging for students - Stimulates creativity and originality - Efficient with smaller student numbers	- Labour-intensive checking - Students may process feedback too superficially - Requires the language skills of the student to be sufficient	- Cognitive skills at the knowledge, understanding, application, analysis, synthesis and evaluation level.
Papers	- Challenging for students - Stimulates creativity and originality - Suitable for integrative testing of multiple skills	- Reliability of assessment requires additional care - Labour-intensive checking - Student may process feedback too superficially - Tests skills which may not be learning objectives (language skills, organisational skills)	- Higher cognitive skills such as application, analysis, synthesis and evaluation - Written communication skills, such as writing a research report, paper or essay.
Assessments	- Better reflection of later professional practice	- Reliability of assessment requires additional care - Time consuming, in regard to both organisation and administering the exam	- Application and integration of skills which cannot be assessed in written exams.

Open questions. With an open question, whether or not introduced with contextual information, the student must substantiate their answers. Exams with open questions stimulate students' creativity, reasoning skills and independent thought which is characteristic of the academic outlook. Exams with open questions are therefore preferable to those with multiple choice questions.
The disadvantage of using open questions is the intensive checking work required. Because of this, open questions are best suited to smaller numbers of student. With larger numbers of students, combinations of multiple choice and open questions could be a solution. In which case, the multiple choice questions would be checked first so that if it becomes clear that a student will fail the exam, the open questions need not be checked.
An additional drawback of exams with open questions is that they also rely on students' language and reasoning skills, even though this is not necessarily a specific learning objective of the course. The disadvantage is relative, however, given that language and reasoning skills for academic students are required in any case. However, it is advisable to also incorporate these skills into the learning objectives.

Papers. Exams with academic papers require students, whether or not in groups, to prepare a text to be assessed. Examples include papers, essays and research reports.

Papers are particularly suited to testing for learning objectives where the application and integration of various spheres of knowledge is central. Thus, students must be able to select, combine and apply this knowledge and must possess sufficient writing and reasoning skills.
A potential disadvantage of using papers to assess students is the relatively low level of reliability of the assessments and the potential for variation between different assessors. This can be reduced (but not entirely ruled out) by the use of rigorous assessment schemes. Excessive detail should be avoided, however.
Furthermore, additional measures must be taken to also assess students individually in situations when a paper is written as a group. A final disadvantage for lecturers is that papers are often labour-intensive to check.

Assessments. Actions in a simulated professional situation are tested and evaluated with an assessment. Examples of this exam type are interviewing, giving presentations and using certain computer programs. This exam type will be used primarily to test learning objectives which students need to have acquired, but which cannot be demonstrated in a written exam. An example of such a learning objective is ‘you can give a presentation’.
A potential disadvantage of assessments is the relatively low level of reliability. In addition, an assessment is a time consuming task that also requires the necessary extra organisation and facilities. For some learning objectives however, assessments are the only possible exam type.

2.3 Guidelines regarding exam construction and assessment for each exam type

This section describes how best to meet the validity and reliability requirements for each exam type and what exactly is involved in the assessment. The following topics will be discussed: multiple choice questions (2.3.1), open questions (2.3.2), papers (2.3.3), assessments (2.3.4) and internships and theses (2.3.5). We use assessments here as a collective term for examination using more or less authentic professional tasks such as presentations, interviews, assessment sessions, conversation skills, statistical operations, formal modelling, software design or evaluation, etc.
In order to provide clarity in the outline, we first provide an overview (Table 5) of the guidelines for each exam type which must be taken into consideration to ensure validity and reliability.

Table 6: *Overview of the guidelines for each exam type*

Exam type	Validity	Reliability	Assessment
Multiple choice questions	- Exam matrix - Plausible distracters - Proper construction - Peer review, both before and after - Psychometric data analysis	- Sufficient number of items - Analysis of psychometric data	- Absolute assessment or absolute assessment with a relative component
Open questions	- Exam matrix - Proper construction	- Sufficient number of questions - Model for checking - Assessment methods, which minimise assessor bias	- Assessment scheme - Second assessor; anonymous assessment
Papers	- Exam matrix - Clear task - Peer review	- Assessment scheme: criteria and procedures - Peer review	- Second assessor; anonymous assessment
Assessments	- Exam matrix - Clear task - Peer review	- Assessment scheme: criteria and procedures - Peer review	- Second assessor - Practicing with assessment scheme - Recording the assessment if necessary

2.3.1 Exams with multiple choice questions

Amount of questions

The minimum required number of questions in a multiple choice exam is first determined by the learning objectives: each learning objective must be tested at least once. The more weight a learning objective has, the more often questions will be asked about it. Also, the more questions there are in an exam, the better students can demonstrate that they have mastered the learning objectives. The exam is then a better sample of the study sphere. Table 7 shows the minimum number of questions required per number of answer alternatives in order to ensure a reliable exam[17].

Table 7: *Minimum number of questions per number of answer alternatives.*

Number of answer alternatives	*Number of questions*
Four alternatives	40
Three alternatives	60
Two alternatives	80

Number of answer alternatives

The quality of the multiple choice questions depends heavily on the quality of the distracter choices. The recommendation is to make the decision of choosing between four

[17] Berkel, H. van, Bakx, A. & Joosten-Tenbrinke, D. (2013). Toetsen in het hoger onderwijs (Third edition). Houten: Bohn, Stafleu, van Loghum.

or three answer alternatives based on the number of good distracter choices available, partly in relation to the time investment. Lecturers often seem to choose four answer alternatives with the idea that the chance of guessing correctly will be lower. However, the rate of guessing is often higher than assumed, given that the quality of the distracter choices goes down with four answer alternatives[18].
It is possible to use two, three and four answer alternatives in one exam. It is a good idea to disclose this to students in advance. When checking the different probabilities of guessing should be accounted for. These exams can simply be analysed by the Integrated Accessibility Standard, provided that the lecturer indicates in the answer key how many answer alternatives each question has.

Careful formulation of alternatives
Ensure that the answer alternatives are formulated well:

1. The answer alternatives are all focused on the same aspect.
2. The answer alternatives are all about the same length.
3. The answer alternatives are equally nuanced.
   Example where Alternative A is more nuanced than B:
   A) the group of cells which is activated during the development of memory
   B) neurons
4. Arrange the answer alternatives neutrally, so that the order does not provide implicit clues which lead to the correct answer.
5. Ask one question at a time and also only provide one answer in each answer alternative.
6. Do not give unnecessary information in the answer alternatives; remove data which is not distinctive.
7. The answer alternatives should be mutually exclusive. Avoid overlap.
   Example of overlap:
   A) people with psychiatric disorders
   B) people with an Axis-II disorder
   C) people with a narcissistic personality disorder
8. An answer alternative should be defensibly correct, the others defensibly incorrect.
9. Avoid logical or content clues.
   Example in which the correct answer can be deduced from the sentence structure:
   'Once a thief, always a thief.' What does this saying mean?
   A) If you cheat someone, you will also be cheated.
   B) If you are dishonest, you can never forget that.
   C) If you are bad, you will also think ill of others.
   D) Whoever transgresses once, will always be untrustworthy.
10. Avoid absolutes (never, always) and wording that is too open (can, sometimes, perhaps) in the answer alternatives.
11. Keep it as simple as possible and avoid complicated or metaphorical language.

[18] E.C. Paes, E.C. & Cate, O. ten. (2009*).* Meerkeuzevragen met drie, vier of vijf alternatieven: wat is beter? *Tijdschrift voor Medisch Onderwijs,* 28*(3).*

True / false questions
The same guidelines that apply to multiple choice questions also apply to true / false questions or statements. There are some additional recommendations: ensure that the entire statement is correct and that the formulation is accurate. The student must, after all, judge an entire proposition as true or false, and thus, there can be no doubt about any part of the formulation. Additionally, consider limiting the difficulty of a proposition by giving instructions to only assess the correctness of a specific word or phrase set in italics.
Validity is compromised with answer alternatives in the format below.

A. Only 1 is correct
B. Only 2 are correct
C. 1 and 2 are both correct
D. 1 and 2 are both false.

These are actually two true / false questions forced into a four answer alternative. This type of question is not recommended. First, because it is unfair to students: if they correctly indicate that 1 is correct, but have no idea about 2, they get no points for 1. Second, a greater number of questions increase the reliability of the exam and thus, splitting the questions is desirable. Third, the feedback to the student is less accurate: if the student answered the question wrong, we still do not know what the student possibly knows at the level of each individual statement. For these reasons, it is better to split such questions into two separate true / false questions.

Checking exams with multiple choice questions
Students' raw score on the exam with multiple choice questions must be converted into a grade. It is important to, first, determine the grading factor: so what score will earn a 5.6 (to be rounded up to 6) and thus a passing grade, and what score will earn a 5.4 and with it, a failing grade. Based on this, the further range of numbers can be determined.
Two methods can be used here: the absolute assessment or the relative assessment. Both methods have their proponents and opponents. With the *relative assessment*, a fixed spread is maintained, regardless of the level of the group of students. For example: the top 30% receive 7 or higher; the lowest 30% fail; and the middle group passes with 6 or 7. With *absolute assessment*, the grading factor is used regardless of the actual level of the group of students. The School's examination guidelines are intended to be used with *absolute assessment*. This is possible when multiple choice exams are composed on the basis of an ample sampling from an exam database, which is psychometrically monitored. In certain cases, absolute assessment is not a good option. For example, when new exam questions are designed or psychometric data is otherwise lacking. In such instances, the exam developer has no control over the difficulty level of the final exam. As a result, it is possible that one exam is significantly more difficult than another. An absolute grading factor is then not applicable, since the pass criteria will then be unequal between different groups of students. In that case, the examination guideline is used to find a compromise between the two, which is the *absolute method with a relative component*. By this we mean a method in which the grading factor is related to the average score of the top 5%.

The method of calculation of 'absolute assessment' and 'absolute assessment with a relative component' is described in Appendix II.
A second reason for choosing the *absolute method with a relative component* is that the maximum score for a high-performing student on a multiple choice exam is lower than the theoretical maximum score: as a result of the correction for guessing, a student can never achieve the maximum score (a grade of 10), even if he/she may have answered all the questions correctly. By basing the scoring on the top 5% of students, the highest achievable score is theoretically possible.
Absolute assessment with a relative component is only justifiable if the groups of students are large (larger than 400)[19], they are comparable from year to year and the quality of education has remained the same. Research shows that when the difficulty level is corrected for, the pass rate fluctuates less over the different years and a larger percentage receives passing grades[20]. It also appears, from the same study, that there is no knowledge loss with exams where absolute assessment with a relative component is applied: students seem to achieve the same level of learning as students in programmes that grade exams using absolute assessment.

Maintaining standards for resits
Since the number of students requiring a resit is not the same as at the first exam sitting (the resit group mainly includes the relatively weak students and a smaller group composed of students who failed based on percentages), it is undesirable in principle to make the grading factor dependent on the performance of the group. That would effectively mean that the resit would probably be assessed 'more flexibly' than the initial exam. This increases the likelihood that students will receive an unjustified passing grade. On the other hand, the difficulty level of resits can also vary. In order to cope with this problem, at the School, we have chosen to go with the same level of difficulty with the resits as the initial exam and thus maintain the same standard[21]. The condition is that the resit is composed in the same manner as the initial exam. This means that in the case of the 'absolute assessment with a relative component', the average score of the top 5% of the resit is taken as the average score of the top 5% in the initial exam.

2.3.2 Exams with open questions

Just as with multiple choice questions, it is also true for open questions that they must be related to the learning objectives to ensure the validity of the exam. The exam matrix (see 2.2) is also an excellent tool here.
The *validity* of open questions is also largely determined by the quality of the question formulation and should take the following points into account:

1. A problem which can occur with open questions is *ambiguity*. Essentially, this means the question has different answers, all of which are defensible. To prevent

[19] Sanders, P. (2011). *Toetsen op school.* Arnhem: Cito.
[20]

[21] Gruijter, D.N.M. de (2008). *Toetsing en Toetsanalyse*. Leiden: Leiden University.

this, it may be a good idea to work inversely: first formulate the desired model answer and then the corresponding question.

2. It may also be necessary to add answer limitations to the question. For example, students are often more succinct or comprehensive in their answers than the lecturer intended. Information about the desired length of the answer can be helpful here. Be specific. Formulations such as 'name some examples of …' are too vague. A pre-structured area for filling in the answer can also be useful.
3. Make sure that it is clear which part of the question the answer must cover. A question like 'explain why during the treatment of little Hans, Freud sat behind his patient' for example, can be read with the emphasis on *why*, on *Freud*, on *little Hans* or on *behind*. Depending on this emphasis, the answers are likely to differ. This can be solved by, for example, underlining or italicising the part with the correct emphasis.
4. If you want students to motivate their answer, state this clearly in the notes to the question. Similarly, indicate that students' knowledge from the subject matter x and y should be used when answering.
5. Ensure that the question is linguistically correct and formulated as simply as possible. Avoid e-language.
6. Formulate the question positively (say what the student must do, not what they must not do).
7. Check whether the question contains sufficient information to be able to provide an optimal answer and whether redundant information has been removed.
8. Do not ask trick questions.
9. Provide a clear layout.

To ensure the *reliability* of an exam with open questions it is critical, just as with multiple choice questions, to ask a sufficient number of questions so that all learning objectives are covered and each learning objective is addressed in several ways if possible.
In addition, the *assessment criteria* are important for the reliability of the exam. With open questions it should be clearly stated, prior to correcting, which criteria the assessment will be based on and what particular performance leads to what assessment. This increases the objectivity of the assessment. The following guidelines ensure the quality of the assessment criteria:

1. Formulate a correction model at the same time as the exam construction. Exactly what performance from the student will earn what assessment must be clearly and carefully established. This is often done by formulating a *model answer* for each question. The correction model is primarily intended for the assessors and to substantiate the grades during the review.
2. A good correction model goes beyond just a model answer. The correction model includes the distribution of points linked to the answers and, for instance, also provides examples of clearly incorrect answers.
3. Also clearly state general assessment instructions in the correction model. For example, how to deal with partially correct answers, mistakes which impact following questions, language and spelling errors, illegible handwriting and exceeding the prescribed maximum word limit for answers.

4. Test the correction model in advance for usability and completeness. You can do this by having someone take the exam who has mastered the material, but for whom the exam itself is new. The assessors can also do an interim evaluation of the correction model once about a third of the exam has been corrected. Is the correction model complete? Are all criteria usable?

The reliability of an exam with open questions can also be increased by choosing a useful assessment methodology and by reducing any effects of the assessors which may occur. Table 8 describes common assessor effects and offers advice for minor adjustments in the assessment methodology to counteract these effects, whenever possible.

Table 8: *Common assessor effects and measures to counteract them*

Assessor effect	Measures
Shift in standards: During the assessment, the assessor becomes increasingly stricter or more lenient.	Vary the correction sequence of exam components. Correcting alphabetically is strongly discouraged.
Sequence effect: After numerous poor performances, the assessor may award a relatively good performance a disproportionately high grade (and vice versa).	Correct open questions preferably per question and not per student. In this way, the sequence effect and usually also the halo and contamination effect is reduced. Additionally, this also usually saves time as the assessor does not always have to continually switch between questions.
Halo effect: The image that the assessor has of the student influences the assessment (for example, 'he/she is a good student').	The exam is corrected by an assessor who does not know the student. Assess anonymously (using student numbers)
Contamination effect: The assessor views the performance of the students as a reflection of the quality of his/her own teaching and is inclined to assess students' performances higher than is realistic.	Disconnect teaching and assessment as much as possible.

2.3.3 Exams with papers

As with other types of examinations, an exam matrix can also be created for assessing papers. Instead of the exam questions, in this case, the assessment criteria from the assessment overview in the matrix are linked with the learning objectives associated with the assignment.

To increase the validity, reliability and transparency, it is crucial to provide a specific description of the expected final product (*What* is being assessed?), the assessment criteria (*On which* criteria is the assessment based?), and the grading (*How* is it weighted and graded?). Thereby, students have sufficient information for the execution of the assignment (preparing the paper) and the lecturer has sufficient information in the

assessment of the assignment. The entire set of assessment criteria and assessment procedures is called the assessment overview.

In exams with papers, as with other exam types, the assignment and the assessment overview should be tested, preferably in advance. For example, the assignment and the correction model can be evaluated through peer review. The following questions should be discussed and agreed upon in this evaluation:

1. Has the assignment been clearly formulated? Is the assignment unambiguous? Is the assignment correctly formulated linguistically?
2. Does the assignment include all information students need to complete it optimally? Does the assignment contain any superfluous information?
3. Is the correction model unambiguous?
4. Do the assignment and the assessment criteria link well with the associated learning objectives?
5. Is there mutual agreement on how the use of the model leads to a final grade?

Distinguishing between supervision and assessment

In practice, the supervisor for an assignment (for instance, the thesis supervisor) often acts as an assessor as well. However, this dual role is accompanied by the risk that the supervisor may allow his/her view of the student and the relationship he/she has established to influence the assessment (halo effect). Another risk is that he/she may assess himself/herself (contamination effect). These effects are magnified if not only the product is assessed, but also the process.

In order to counteract these effects, it is recommended that the paper is assessed by someone other than the supervisor. With internships and theses, an independent assessment by a second assessor is mandatory (blind four eyes principle). If these individuals' assessments differ significantly, the procedure requires that both assessors consult with each other to reach a unanimous assessment, based on the agreed criteria. If this is unsuccessful, then the matter will be transferred to the coordinator of the course concerned, who will then determine the final grade.

Individual assessment of group processes and products

For papers that have been produced by a group of students, it is more difficult to award individual grades. However, as this is still desired, the lecturer, for example, can systematically record his/her observations, can have students keep a logbook of activities, questions and ideas, or can have students do individual presentations or write reflective reports. Students can also write portions of the paper individually and a portion as a group.

In group work, it is important to separate the group process from the group product – the paper. The group process is only assessed if 'collaboration' is part of the learning objective. If that is the case, then it is important to explicitly state the corresponding assessment criteria early on, and with this, also devote interim attention to this by providing feedback from the lecturer or *peer feedback* from the members of the group. In addition, the lecturer can also choose to use *peer assessment* (students assess each other) for the assessment. Students should be thoroughly prepared for this.

2.3.4 Exams with assessments

Examination on the basis of assessments (more or less authentic professional situations) generally corresponds to exams with papers, as described above. An important difference however is that an assessment is evaluated on site, while a paper can be assessed at the pace and time the assessor so wishes. For this reason, it is crucial to have a strong assessment overview and for the assessors to be practiced in its use. Recording the assessment in picture and/or audio form is a helpful tool. Thus, the assessor is able to replay the video and/or audio if that is necessary for the evaluation. This also helps when giving feedback to the student.

2.4 Guidelines for exam evaluation

So far we have described the methods which should be used to guarantee exam quality *before* the exam sitting. Psychometric analysis *after* the exam sitting provides a great deal of information regarding the quality of the exam questions and of the exam as a whole, both in terms of validity and reliability. This information therefore contributes to the quality of the students' performance evaluation and provides further clues for improving the exam quality.
Multiple choice exams are psychometrically analysed by the Institute for Applied Social Sciences (ITS). In principle, all exam types are suited to psychometric analysis and interpretation. At the moment however, no examples of psychometric analysis of other exam types are available.

The psychometric analysis of multiple choice exams is explained below, with regard to validity (2.4.1) and reliability (2.4.2). We also discuss the question of how psychometric values can be interpreted and what decisions the examiner may consider (2.4.3).

2.4.1 Validity

F value
The *F value* (frequency) per answer alternative indicates the absolute number of students which chose the particular answer alternative. Thus, it is a measure of the quality of the distracters. Ideally, the students who answered a given question incorrectly should be equally distributed over the distracters. This shows that all distracters have the same opportunity to be chosen by someone who has not mastered the material.
In practice however, we often see that one or more alternatives are rarely or even never selected. In that case, it seems that even without much knowledge of the material, students know that this alternative is definitely not correct. The question thus does not only measure the extent to which the student has mastered the material, but apparently also something else. The validity thus decreases.
If distracters are found to be selected too infrequently in an exam, then they must be removed. Simply use the question with one distracter less, of course, don't forget to adjust the question for the chance of guessing. Creating a new distracter is of course also a possibility.

P value

The *P value* is an indication of the degree of difficulty per exam item and provides information about the selective capacity of the exam. The *P value* is determined by the proportion of students who answered the questions correctly and is therefore always a number between 0 and 1. Typically, two different P values are given: the *p* and the *p'*. Both show the proportion of students who answered the question accurately, whereas *p'* is adjusted for guessing. A *p* of 0.80 indicates that 80% of the students answered the question correctly. A percentage of them will have guessed the answer however. The *p'* is adjusted for guessing. In the case of a four-choice question this value is then 0.73. (see: Appendix II). The question can now be seen as a question which was answered accurately by 73% of the students, without needing to guess.

By including exam items of varying degrees of difficulty in an exam, it is possible to differentiate between the different performance levels of students. The *p'* values show to what extent this was successful. A question with a *p'* value of 1 was answered correctly by everyone and therefore, this question has little to no capacity for selection. The same applies to a question with a *p'* value of 0, in the sense that no one answered the question correctly. A *p'* of 0.5 seems desirable as it provides the maximum contribution to the selective (summative) function of the exam.

To correctly interpret the *p'* value, it is important to take the nature of the group of students who has answered the question into account. Since the *p'* value is group-dependent, this value will usually be lower in groups with a generally lower level (such as a resit group) or smaller groups in which chance plays a more significant role.

2.4.2 Reliability

Coefficient alpha

The *coefficient alpha* is the measure of internal consistency, and thus indicates the reliability of the exam as a whole. The internal consistency of the exam shows the degree to which the exam items are statistically coherent with each other. Coefficient alpha always lies between 0 and 1 and the higher the better (0.7 and above is generally acceptable). A low coefficient alpha can be increased next time by including more questions with a higher R_{ir} (see below) in the exam. Exams that have already been taken and wherein the coefficient alpha was found to be too low, the decision can be made to remove questions with a negative R_{ir} from the exam and thus increase the coefficient. If this means that a large number of questions must be removed from the exam, this means the exam will obviously be less representative of the learning objectives and a new exam will need to be compiled for the next sitting.

Item residual correlation

The *item residual correlation* (R_{ir} value) is a measure of the discriminating power of an exam item. Each exam question must ideally make the best possible distinction between students with a high and a low grade. The R_{ir} can be used to check whether the exam questions comply with this criterion. The R_{ir} values of the individual items are related to the reliability of the exam as a whole. The R_{ir} is determined by linking the score on the item to the final score of the entire exam (adjusted for the score on the item concerned) and consists of a number between -1 and 1. A strongly positive R_{ir} indicates that students

who answered this question accurately also had a higher average score on the exam than students who answered it incorrectly. The R_{ir} must, in any case, be positive with a target value of > 0.2. If the R_{ir} is negative or very negative, then the question does have discriminatory power, but the wrong discriminatory power. Roughly speaking, it means that the question was answered accurately by the poor students and inaccurately by the good students.

2.4.3 Exam quality improvement

Based on the quality indicators that are described above, it is possible to improve the quality of the exam *after* the exam sitting (Table 9). This requires an accurate interpretation of the psychometric analyses.

Table 9: *Interpretation of possible combinations of p' and R_{ir} values and desired measures*

	R_{ir} is negative	**R_{ir} is lower than 0.15**	**R_{ir} is higher than 0.15**
***p'* lower than 0.1**	This is a poor question: Is the answer key correct? *> remove the question from the exam if the answer key is correct.*	This is a question that seems to have something wrong with it: Is the answer key correct? Is it a detailed question that does not correspond with the learning objectives? Is the formulation of the question unambiguous? Is another alternative plausible? *> remove the question or allow for multiple possible answer alternatives, if warranted.*	This is a difficult question, and only distinguishes between the nines and tens: Trick question? Too difficult/complex? *> keep the question, just make sure that there are not too many of this type of question.*
***p'* between 0.1 and 0.8**	This is a poor question: Is the answer key correct? *> remove the question from the exam if the answer key is correct.*	This question is not too difficult or too easy, however, there could be another reason why the question is not discriminatory. Is another alternative also plausible? *> possibly allow for multiple answer alternatives.*	Excellent question. *> keep the question*
***p'* higher than 0.8**	This is a poor question: Is the answer key correct? *> remove the question from the exam if the answer key is correct.*	This question is too easy and doesn't discriminate. Is it a giveaway (can be solved with common sense)? Are the distracters sufficiently plausible? *> does not require immediate action, but should be modified in the database.*	This is an easy question and only distinguishes between the ones and twos. *> keep the question, just make sure that there are not too many of this type of question.*

However, there are a number of conditions which must be taken into account when adjusting the exam based on the psychometric data:

1. It is important to always interpret the quality of a question or an entire exam in combination with the content of the exam items themselves and never solely on the basis of psychometric data. This type of quantitative analysis always involves margins based on chance and thus should not be taken as absolute.

2. Take the number of exam participants into account. The fewer participants, the less reliable the quality indicators are. This particularly plays a role in re-sits. In this particular case, the sampling is a non-heterogeneous group, and any decisions based on the psychometric data must therefore be made with extreme caution.

3. When making decisions regarding adjusting the exam, it is crucial to interpret the *p* and R_{ir} values in combination with each other. Table 9 provides an overview of all the possible combinations and the corresponding recommended measures. An explanation of the most important combinations is given here:

Combination of a very low p' value and a low R_{ir}: this combination is suspect. Always check the contents and answer key to such questions for accuracy and remove the question if there is cause for concern.
Combination of a low p' value with a negative R_{ir}: this combination is a reason to remove the question from the exam (assuming that the answer key has been used correctly). First of all, the question is too difficult as it only discriminates between the poor students and therefore, not between the good students and the poor performing students. Second, the good students score worse on this question than the poor students, which is exactly what you want to avoid.

Combination of a low p' value in conjunction with a sufficiently positive R_{ir}: such a question may be kept. However, make sure that there aren't too many of these difficult questions in the exam, because, ultimately, you especially want to distinguish the fives from the sixes.

Remove poor questions or properly allow for all possible answer alternatives?
Another alternative to removing the question from the exam is to allow for multiple correct answer alternatives. It's best to opt for the latter when the content of the question and the answer alternatives justify this, for example, if one of the distracters proves to not be entirely wrong upon closer inspection. If it is decided to remove a question from the exam, keep in mind that the reliability of the entire exam can be negatively affected with fewer questions. You must ensure that the exam is still sufficiently representative of the learning objectives of the course.

Appendix I: Setting learning objectives (or improving them)

Learning objectives are:

- explicit (unambiguous and specific). *See below under: SMART, RUMBA*
- limited in number and ordered hierarchically: the most important objective is named first; consider using sub-objectives (e.g. 2a, 2b)
- at the level appropriate to the programme phase and within the study track
- formulated in behavioural terms (actions) or cognitive performance
- still relevant after the studies, so relevant to the future life of the students as professionals, academics and productive citizens.

1. Editorial tips for changing existing course objectives

- Begin with: On completion of the course
- Use *you* instead of *the student*.
- Use active verbs (see below): you analyse, you substantiate, you compare, you assess, etc.

2. Do you currently have too many course objectives?

- Reduce the number of main objectives to about 3 to 5
- Name the most important objective first; in the former layout, it often came last.
- Some of the objectives are likely conditional on the ability to achieve the main objectives. Formulate these conditional objectives as sub-objectives (e.g. as 2a, 2b). This creates a clear 'hierarchy of objectives' which helps the students, lecturers and assessors to work with more of a focus on the learning objectives.
- Now consider whether the objectives and the order are really crucial (in this course and at this level)?

3. Do you currently have course objectives that start with 'Knowledge and insight in …'?

- Knowledge and understanding are certainly needed, however, be specific about what kind of knowledge this course requires, and what type it does not require.
  Not this: knowledge from subject A
  But this: knowledge of three common personality theories, namely X,Y and Z
- Ask yourself what kind of cognitive performance or what type of academic behaviour you expect from a student who possesses this knowledge. What should a student do/demonstrate to convince you that he/she possesses this knowledge and knows how to utilise it?
  Your answer to these questions will probably contain a specific course objective, in which knowledge and understanding (*Knowledge* and *Understanding* level by Bloom) is conditional in order to achieve a higher cognitive level. Use, for example, one of the verb forms of 'analysis', 'evaluation' or 'synthesis'.
  See below under: Notes on academic level.

4. Do you currently have course objectives containing 'identify', 'recognise', 'compare' and 'describe'?

- These verbs mean that your course objectives are at the *Knowledge* and *Understanding* level. This is an appropriate level for introductory courses.
- Such a level is often not sufficient for subsequent courses. To resolve this, it is useful to try to complete the following sentence for yourself: 'The student must identify/recognise/compare/describe so that he/she can then X and Y'. Completing this sentence will probably lead you to a course objective at a higher level; especially if you use one of the verb forms of 'analysis', 'evaluation' or 'synthesis' in the sentence.
  See below under: Notes on academic level.

Specifically worded using SMART or RUMBA

To make your course objectives more specific, consider the 'SMART' criteria:

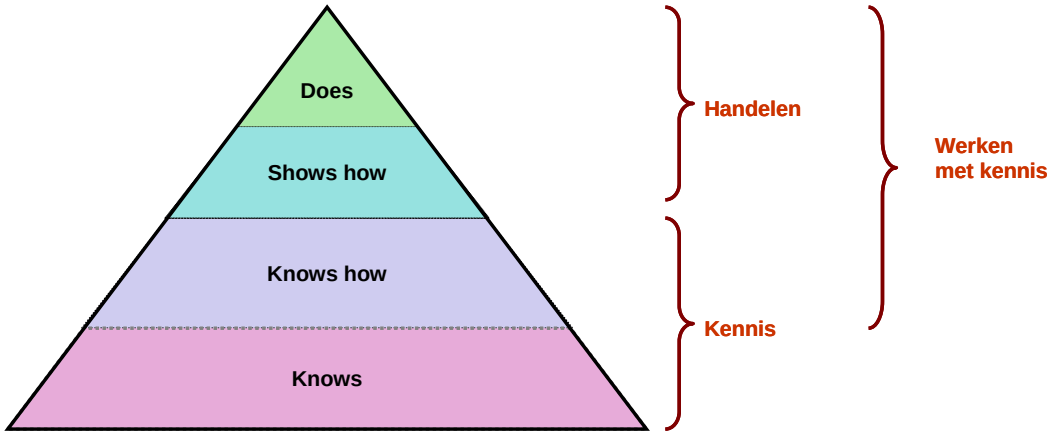
- Specific
  Is the formulation in specific and understandable terms? Is the context clear?
- Measurable
  If the objective is achieved, the result must be measurable. Is that reflected in the formulation of the objective? Do students have, for example, an idea of what content will be asked of them in the exam?
- Attainable
  The formulation of the objectives should be recognisable to students as meaningful and relevant in view of their own learning needs in relation to the programme.
- Realistic
  Have the objectives been formulated in such a way that they can be viewed as achievable, given the level or prior knowledge of the student?
- Time-bound
  Are the objectives achievable within the time reserved for them? But also: does the course come at the right time given the curriculum?

There are several lists, such as RUMBA: *relevant, understandable, measurable, and behavioural.* All lead to similar results as they help you find the balance between learning objectives which are too abstract, incomprehensible or boring on the one hand, and learning objectives which are too common or too application-oriented on the other. They also help you to determine whether the objectives are at the right level.

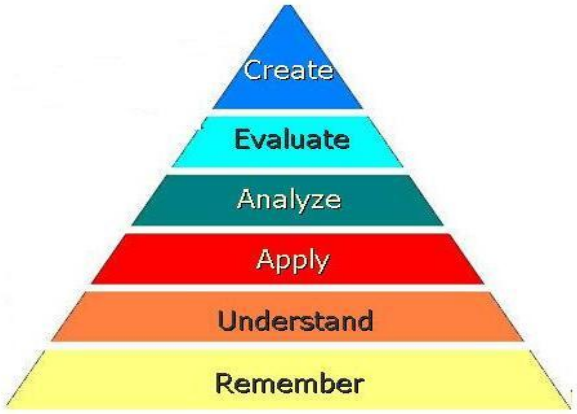
Academic level

Higher education is aimed at allowing students to learn in a meaning-oriented or application-oriented manner, rather than reproduction-oriented or unfocused. Lecturers should allow students to *engage with knowledge*, in other words, lecturers should focus their course objectives on the 'higher order thinking' skills (metacognitive skills).
We use two tools to formulate course objectives at the appropriate academic level: the Miller pyramid (1990) and the (revised) Taxonomy by Bloom (Anderson & Krathwohl, 2001). These are shown below.

Academic course objectives should (at least) be incorporated at the **'werken met kennis' (working with knowledge)** level or at the level of **Application and higher** (Apply, Analyse, Evaluate and Create) (Anderson & Krathwohl).
Typical verbs belong in the different cognitive levels (see below). These can be used to actively formulate your learning objectives to ensure the student does something.



Miller Pyramid (1990)



Revised Bloom Taxonomy, Anderson & Krathwohl, 2001

Apply

Conducting or using a procedure, model or theory for specific situations and problems which are new to students. Characteristic questions: How does that work here? What is needed for this?
Characteristic *verbs* are: choose, demonstrate, construct, conduct, predict, translate, use, execute, implement, etc.

Analyse

Splitting a larger whole into its constituent parts, figuring out and distinguishing step by step the different aspects of a problem, thought or theory. Characteristic question: How does that work? What parts are of greater/lesser importance? How can they be ordered? Characteristic *verbs* are: select, compare, contrast, investigate, categorise, classify, distinguish, etc.

Evaluate

Forming an assessment based on criteria and standards/methods by checking facts and investigating and critiquing assumptions.
Characteristic *verbs* are: assess, test, critique, support, defend, substantiate, etc.

Synthesise

Critically considering or developing something new. Contributing ideas together with authors, lecturers and fellow students. Bringing own ideas to the table and not simply accepting everything that is written or said.
Characteristic *verbs* are: combine, reformulate, summarise integrally, argue, infer, generalise, conclude, criticise, problem solve, innovate, decide, recommend, etc.

Appendix II: Methods of calculation: 'Absolute assessment' and 'absolute assessment with a relative component'

Absolute assessment

In order to receive a passing grade, the student must answer more than half of the questions correctly, after correcting for the chance of guessing. The formula for the absolute standard is therefore:

$$((5.6 - 1) * (M - T) /9) + T = \text{number of questions correct for a pass}$$

Whereby:
5.6 – 1 = Passing grade minus 1 point since it is impossible to receive a grade lower than 1.
M = Maximum achievable score
T = Total number of questions / number of answer alternatives (= expected number of correct answers based on the chance of guessing)
9 = This is the range in which you want to award points (we calculate from 1-10).

Sample calculation:
Suppose that an examination consists of 60 four-choice questions. The chance of guessing is 0.25, so based on the chance of guessing, an average of 15 questions will be answered correctly[22]. The next score is then graded at 5.6 and after rounding up this becomes a 6:

$$((5.6 - 1) * (60-15)/9) + 15 = 38$$

Absolute assessment with a relative component

An absolute assessment with a relative component is not based on a theoretical maximum score (all questions correct), but rather the maximum score that has been shown to be 'achievable' in practice. Because of this, the following definition is used in the literature: the average score of the best 5% of the corresponding exam. The following formula shows what score leads to a pass (5.6):
In the above formula, the practically achievable maximum score is used instead of the theoretical maximum score (M).

Sample calculation:
Suppose that an examination consists of 60 four-choice questions. You can see the average score of the best 5% of the students in the frequency distribution of the exam scores, for instance 58. The chance of guessing is 0.25, so based on this an average of 15 questions will be answered correctly. The next score is then rated at a 6:

$$((5.6 - 1) * (58-15)/9) + 15 = 37$$

[22] Van Berkel et al., 2013.

Awarding grades

Using the same formula, it is also possible to award all raw scores a grade between 1 and 10. The adjusted formula that can be used for this purpose is as follows:

$$\text{Grade} = \frac{(X-A)}{(T-A)/9} + 1$$

Whereby:
T = Average score of the top 5% (using a relative component) OR the Theoretical maximum score (all questions correct) (using absolute assessment)
A = Total number of questions / number of answer alternatives
X = Number of questions answered correctly (raw score).

Sample calculation:
In an examination with 60 four-choice questions, a student answered 37 correctly.
T = 58
A is 60/4 = 15
X = 37

The grade is then:

$$\frac{(37 - 15)}{(58 - 15)/9} + 1 = 5.6$$

By putting this formula into SPSS or Excel, all the students' raw scores are turned into grades between 1 and 10. An Excel program for this is available from the Quality Assurance Team.