

Guidelines on Grading Written Examinations

A PDF version of these Guidelines, with links, is available electronically at www.let.ethz.ch/docs/Guidelines_GradingEN_2013_11.pdf

A checklist for implementation in practice is found in Appendix 1 and at www.let.ethz.ch/docs/Checklist_GradingEN_2013_11.pdf

Imprint

Issued by: ETH Zurich, Educational Development and Technology

Editing: Tobias Halbherr, Claudia Schlienger

Translation: Katherine Hahn

Printed by: FO-Fotorotar AG

Number of copies: 600

1st edition, November 2013

ETH Zurich

Lehrentwicklung und -technologie [Educational Development and Technology]

Haldenbachstrasse 44

8092 Zurich, Switzerland

www.let.ethz.ch

Table of Contents

Foreword	5
Grading: A real challenge, with measurement theory pitfalls	6
Chapter 1: Principles of grading at ETH	8
Principles governing high-quality written examinations	8
Chapter 2: Good grading practice	10
2.1 Grading scale and allocation of points	10
2.1.1 Determining the grading scale	10
2.1.2 Determining the '4' grade	11
2.1.3 Determining the '6' (and the '1') grade	11
2.1.4 Allocating points: Quantifying examination tasks	12
2.1.5 Points for subtasks	12
2.1.6 Allocating points for multiple choice tasks	13
2.1.7 Scoring chart and sample solution	13
2.2 Correction and scoring	14
2.2.1 Rate the route or the result?	14
2.2.2 Correcting examinations: Eliminating interference	14
2.3 After the examination	15
2.3.1 Viewing the examination	15
2.3.2 Reviewing the examination	15
2.3.3 Identifying problematic questions: Item analysis	16
2.3.4 Dealing with examination errors during correction	16
Chapter 3: Principles	17
3.1 Aligning examinations with teaching and learning activities	17
3.1.1 All of a piece (Alignment)	17
3.1.2 Knew a lot, understood nothing? (Taxonomy of learning goals)	18
3.2 Examinations are measurement instruments	18
3.2.1 When is an examination good? (Psychometric criteria)	18
3.2.2 Examinations measure competences (Explanation of variance)	19
3.2.3 Is a 6 twice as good as a 3? (Scale levels)	20
Appendix	22
Appendix 1: Correction and grading checklist	23
Appendix 2: Further reading	24
Appendix 3: Further resources	24
Appendix 4: Mathematical formulae	25
Appendix 5: Overview of interfering factors	26

Foreword



Maintaining the quality of teaching at ETH Zurich and improving it wherever possible is a major concern of mine – not least with an eye to international competition. Here our examination system and grading are of particular importance. Important teaching quality themes find expression in grading, such as the purpose of examinations and the character of the selections made. Grading indicates which achievements are sufficient, which are very good, and which are not good enough.

As an ETH Zurich lecturer, grading is completely your responsibility. These Guidelines are intended to help you in the task. They do not replace exchange with colleagues. Rather, they firstly show new faculty how to properly take into account grading when designing examinations. Secondly, they address central points useful to all lecturers in the practice of grading.

The Guidelines on Grading have been kept short, but still provide a comprehensive overview of the subject. They illustrate five principles of grading using concrete examples, and the checklist in the Appendices summarises all the important day-to-day procedures. A section on theoretical principles which lists further reading also enables faculty to consider their own practices and learn more about the topic.

I hope that you will do precisely this, and use the Guidelines as an instrument to further improve your teaching.



Prof. Dr. Lino Guzzella, Rector, ETH Zurich

Grading: A real challenge, with measurement theory pitfalls



**Prof. Dr. Elsbeth Stern,
Professor of Research on Learning and Instruction at ETH**

Grades have a considerable impact on educational and career decisions, and may even influence entire life trajectories. From a scientific point of view this is not unproblematic, because grading is a form of performance measurement which is tainted with errors. Generally speaking, a measurement is defined as a set of rules for assigning values to different manifestations of an attribute. This applies as much to the mass and size of an object as to problems solved in an intelligence test, or to grading in classes. Naturally, however, these measurements differ in quality and precision, and measurement theorists have developed specific criteria for these. They are the subject of these Guidelines, together with the question of what conclusions may be drawn from measurements, and which kinds of statistical computations are interpretable. Physical measures such as mass and size have a defined zero point and therefore measurements can be set in relation to one another. Grades, on the other hand, only allow statements regarding ranking. A grade 5 is better than a grade 4, but a 4 is not twice as good as a 2. Not even the distance between grades can be interpreted, because mid-level grades represent a wider spectrum than grades at one extreme or the other. From a measurement theory perspective, arithmetic means of grades may be interpreted as little as may variance. Nevertheless, calculating mean grades is a widely established practice, even though only the median is the permissible as the parameter for the average. With great effort, measurements which go beyond mere determination of rank are also possible in the non-physical area. Measurements of intelligence, personality characteristics or competences are examples. Here every individual examination task undergoes intensive empirical study to ensure that it measures a unified construct. Only tasks of high item-total correlation which differentiate persons of high and low defining (parameter) values from each other are retained. Ideally a professional test follows the so called item-response theory (Rasch Scaling): with high probability persons who have solved a difficult problem have also solved all of the easier problems. In this case the intervals between the measured values may be interpreted. The self-made tests of teachers are a world away from these quality characteristics, and grading is influenced by countless factors unrelated to actual performance. One example is the reference benchmark: frequently the same work is graded more stringently in a stronger class than in a weaker one.

Learning research sees mixed blessings in grading. It often affects the relationship between teachers and learners negatively. Precisely in mathematical and scientific subjects, where deep understanding of concepts is essential,

effective teaching means that errors must not only be tolerated, but must even be welcomed. Instructors need to address the misunderstandings which underpin mistakes and create a climate where learners are not afraid to admit to them. In the graded examination, however, there is no tolerance of mistakes. Moreover, grading may also establish false incentives and priorities, which become clear in the differentiation of mastery and performance goals. A performance goal orientation means getting the best grade with minimum effort instead of focusing on understanding the material. In view of the many problems associated with grading, it is no wonder that correlations between professionally designed performance tests and grades in corresponding subjects rarely exceed $r=.40$.

Despite the shortcomings of grading, educational settings rarely allow the de-coupling of teaching from achievement measurement or delegation of the latter to professional test developers. This may in any case only work if a reliable and valid test which is constantly updated is available in the respective content area. However, in university education this is seldom the case, and because of the large scope of information this will not change any time soon. At universities we must continue to act simultaneously as teachers and as examiners. For this reason it is that much more important to know the pitfalls associated with grading and thus mitigate the problem. Here these Guidelines can be of essential help.

Chapter 1: Principles of grading at ETH

The following principles apply to all forms of performance assessment at ETH Zurich.

- The individual lecturers are responsible for the examination. In particular, they answer for the correctness of content and the methodological appropriateness of examination tasks, scoring and grading.
- Examinations are instruments which measure the achievement of learning objectives or assess the potential to achieve them in the future.
- Examinations adhere to the principles of validity, fairness, transparency, educational purpose and balance described below.
- Design, implementation, scoring and grading of examinations are based upon scientifically grounded methods and standards which are feasible in practice.
- The formal aspects of performance assessments are described in the Guidelines for Lecturers¹. Their legal foundation is the ETH Zurich Ordinance on Course Units and Performance Assessment [Verordnung über Lerneinheiten und Leistungskontrollen (VLK)]² and its implementation stipulations³.

Principles governing high-quality written examinations⁴

1. Validity

The examination is valid, reliable, and tests learning goals objectively. The grade is a meaningful normative estimation of examination performance as a whole.

- Validity:** Examination tasks enable the testing of the competences expressed in the learning objectives in a valid and methodologically acceptable manner. Examination tasks are closely linked to these competences and to the corresponding teaching and learning activities. In content the tasks represent the learning objectives as a whole in an appropriate manner. The grade is a meaningfully weighted value which describes the achievement of learning objectives in their entirety. The cognitive process required to address examination tasks corresponds to the cognitive process associated with the corresponding learning objective. Testing focuses on what is important. Sophistry, complicated and unclear formulations and focus on unimportant details are avoided.
- Reliability:** The examination is sufficiently thorough and differentiates appropriately between various levels of performance. It differentiates results in the critical area (pass/fail) with the highest precision. Repetition examinations are comparable in level of difficulty.
- Objectivity:** Disruptive circumstances during examinations and extraneous influences on scoring do not influence measurement. Exams and scoring take place under uniform conditions. Differences in the evaluation of examination tasks express actual differences in level of student performance and not differing evaluation

¹ See <http://www.ethz.ch/intranet/en/teaching.html>

² See http://www.rechtssammlung.ethz.ch/pdf/322.021_leistungskontrollenverordnung_eth_zuerich.pdf

³ See <http://www.rektorat.ethz.ch/directives>

⁴ These principles also apply to examinations undertaken on the computer.

criteria on the part of individual examiners. Subjective influences on the evaluation of an examination are avoided completely or at least minimised.

- d. Grade scaling:** The grading scale is determined in such a way that a student's grade is not dependent on the achievements of other students. Only pass grades indicate sufficient achievement of learning objectives; the best grade, 6, must be achievable. Equal grade differences (e. g. 3 vs. 4 or 5 vs. 6) reflect comparable differences in achievement of learning objectives.

2. Fairness

Students are never subject to arbitrariness in relation to content, implementation or exam evaluation. All students sit the examination under the same conditions. Equal access to learning infrastructure and learning content is guaranteed. Students' examinations are scored according to consistent and objective criteria.

Examination conditions are conducive to personal best performance. Disruptions, distractions or other interfering factors during the examination are avoided. Only the competences formulated in the learning objectives and their inherent prerequisites have an influence on examination performance and scoring. External factors irrelevant to learning objectives such as sociodemographic characteristics, values or questions of attitude have no influence.

The examination and the examination results must be protected from fraud in their preparation, implementation and post-processing. Examination processing is reliable and error-free.

3. Transparency

Students are aware of an examination's content and formal requirements. This information is easily accessible, complete, understandable and binding. The basis is the entry in the Course Catalogue. Examinations are aligned with the learning objectives communicated: the competences expected from students and tested are concrete, clear and formulated complete, in particular regarding the material's scope and the corresponding cognitive level. The required previous knowledge is also sketched and made known. The examination form and procedure are known. Performance criteria and formal answer structure are determined in advance and communicated. Even before the examination students are enabled an insight into their levels of competence, e. g. via exercises, quizzes or earlier examinations.

4. Educational purpose

Examinations ultimately serve the purpose of improving education. Hence, examination tasks should be aligned with the form, content and demands of the competences targeted and correspond closely to teaching and learning activities. In this they are an incentive to develop the desired competences at the targeted cognitive level during exam preparation. A good examination enables students to demonstrate outstanding skills and motivates them to excel. Examinations are an opportunity for feedback and help to pinpoint strengths or gaps in competences. They are a selection tool and ensure that ETH students fulfil performance requirements. Examinations verify the achievement of learning objectives.

5. Balance

The achievement of learning objectives is tested in a directly plausible and convincingly meaningful manner. The efforts and circumstances of examiners and students stand in reasonable relation to the use or relevance of the examination.

Chapter 2: Good grading practice

The following elucidations regarding grading each aim to satisfy one or more of the principles listed in Chapter 1.

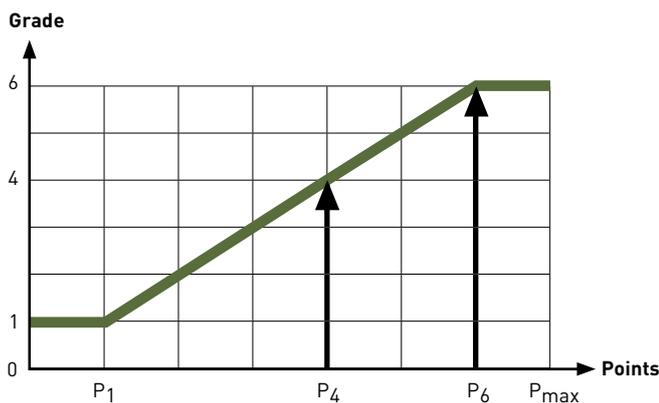
2.1 Grading scale and allocation of points

2.1.1 Determining the grading scale

Examination performance is related to a benchmark via the grading scale. At ETH Zurich grades are criterion-oriented (criterion-referenced).

The mapping of examination points to the grading scale is anchored at grades '4' and '6'. In criterion-oriented examinations performance is mapped against a benchmark that represents a reference level of competence rather than a reference performance from a population of peers. Grade 4 represents marginally sufficient performance and grade 6 outstanding performance. In general all other grades can be satisfactorily computed via a linear interpolation based on these two mappings. Alternatively, fail grades are interpolated separately by defining a required number of points for grade 1. The number of points necessary for grades 4 and 6 are determined before the examination and communicated to the students.

Single linear interpolation



Double linear interpolation

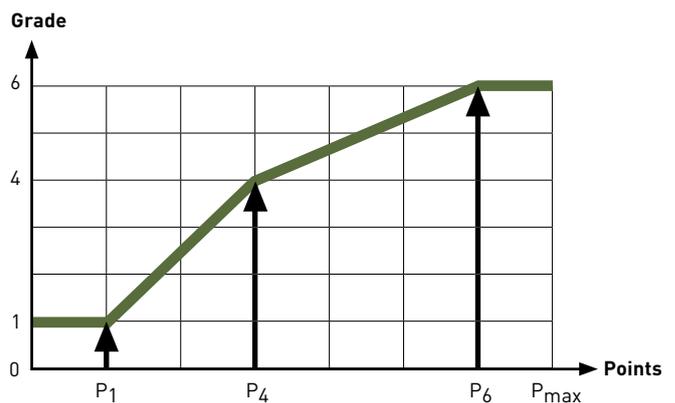


Figure 1: Interpolating grades

2.1.2 Determining the '4' grade

Differentiating between sufficient and insufficient performance is essential. The corresponding number of points must be established before the examination and taken into account as early as the development of exam questions. A grade 4 indicates that the learning objectives have been reached without any margin; its anchoring may not be based on the performance of other students. Where exactly the borderline to a 4 is, however, remains a matter of discretion. Lecturers may use the following criterion-related questions to help them determine it:

- What competences are central to the learning objectives?
- What level of performance reflects minimum sufficient mastery of these core competences, or precludes it?
- What level of performance reflects lasting mastery of competences?
- How well must competences be mastered such that further course units may build upon them?
- What degree of misunderstanding and mistaken or false knowledge precludes sufficient achievement of learning objectives?
- What levels of performance have been considered sufficient up to now?
- What levels of performance are considered sufficient in comparable examinations?

The lecturer should deliberate on what makes for sufficient or insufficient performance even during the question formulation stage, the drawing up of the grading scale and the establishment of sample answers.

The following **methods** help to determine the number of points which equal a 4:

- The examination tasks should be compared on the one hand with the learning objectives and on the other with exercise tasks and earlier examination tasks. The learning objectives specify what is expected of the students. Exercises and earlier exams help to gauge what can be expected of them.
- On the basis of these comparisons an estimate can be made as to the number of points which can only just be considered clearly sufficient (P_g) or insufficient (P_u). The points adding up to grade 4 lie somewhere in between. For 'low-stake' examinations which are not primarily selection-oriented, points are set at $P_u + 1$; in cases of doubt the examination should be passed.

For more precise results several persons should estimate the number of points according to this method, and the number of points should be set individually for sections of the examination.

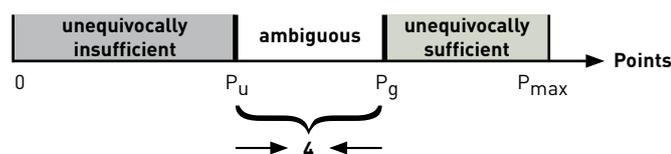


Figure 2 : Estimating the number of points required for 'sufficient' performance

2.1.3 Determining the '6' (and the '1') grade

Goals which are demanding but achievable by one's efforts are strong motivators. Achieving a 6 should be challenging but possible – and at least some students should actually manage it. However, a grade 6 should not simply be awarded to the best students, but should distinguish extraordinary performance. Too easy awarding of good grades devalues them and weakens their positive effect.

If lecturers wish to interpolate fail grades independently of pass grades they must anchor performance to a fail grade, usually a 1. The simplest way is to map zero points to a '1' and interpolate linearly up to a 4. Depending

on the examination content it is also possible to map more than zero points to a '1'. If the examination contains multiple choice questions, chance performance must be taken into account.

2.1.4 Allocating points: Quantifying examination tasks

Examination tasks are quantified in the form of points. The number of points potentially awarded for an examination task may be allocated according to some or all of the following aspects:

- The time required by an expert to complete the tasks
- The length of time required to acquire the competence tested
- The relevance of the learning objective tested

Ideally all three the three aspects correspond to each other: the examination time needed to complete one task corresponds to both the relevance of the tested learning objective and the time expended to reach it. The subjective difficulty of a task, i.e. in the view of individual students, should have no direct influence on the number of points.

2.1.5 Points for subtasks

The total number of points allocated to a task can be differentiated. Here there are three possibilities:

- 1) The task can be divided into **subtasks**.
- 2) The (sub)tasks can be scored according to different **criteria**.
- 3) Varying numbers of **subpoints can be awarded** according to performance per (sub)task on a task subscale.

These three possibilities may be combined (see Fig. 3).

If subtasks are envisioned, note that for each subtask it should be possible to achieve the full number of points, even if other subtasks (and therefore interim results) are incorrect.

Lastly, the individual subtasks or criteria may be weighted by allocating a different number of points for different levels of performance.

Too fine a differentiation between points 1 to 3 is usually counterproductive and can compromise the examination's objectivity. In particular, if a task subscale is deployed a maximum of 5-stage division (0 to 4 subpoints) is recommended.

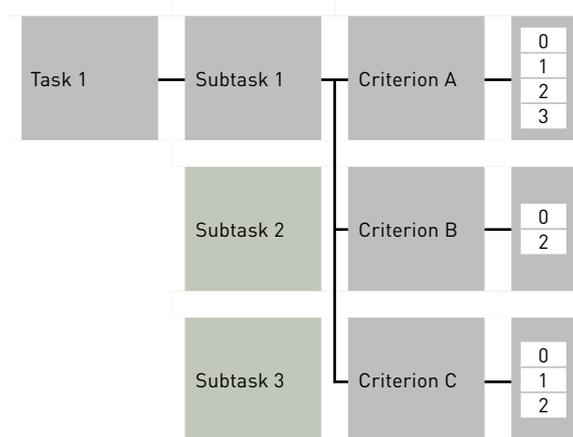


Figure 3 : Differentiating tasks and points

2.1.6 Allocating points for multiple choice tasks

Multiple choice questions take two prevalent forms:

- In **one-best-answer questions** one of (generally) four or five possibilities is unequivocally the right or best answer. The others are unequivocally wrong or worse answers. Points are only awarded for selecting the right/best answers. No points are awarded for selecting the second-best answer and no punitive points are subtracted for wrong answers.
- **True/false questions** are most common in 'K-Prime' (K') format. Here four true/false statements must be considered, of which any number are unequivocally true or unequivocally false. For correct answers to all four subquestions the full number of points is allocated; for three correct answers half the number of points; and otherwise zero points. Alternatively, in true/false questions each correct answer may be awarded one point, and each wrong answer zero points.

The use of any other multiple choice format or scoring method is strongly discouraged. In determining the grading scale the average points obtainable by pure guesswork must be taken into account.

2.1.7 Scoring chart and sample solution

A **scoring chart** defines the formal point allocation scheme and simplifies correction. In the form of a table, it sets out how many points are allocated for which parts of the examination and according to which criteria. In this way the various parts of the examination are weighted in a consistent manner during correction and scored according to consistent criteria. As criteria for allocating points, answers may be approached concretely ('three out of four properties were mentioned') or interpreted qualitatively ('the relevant facts were comprehensively substantiated'). To enable more objective and reliable scoring a good maxim is: 'as concrete as possible, as open as necessary'.

The scoring chart⁵ is drawn up before the examination, discussed with the persons correcting the exam and communicated to the students in an appropriate form. In correcting the examination the reasons for allocating the respective number of points should be added in writing.

⁵ For an example of a scoring chart see: www.let.ethz.ch/docs/Example_ScoringChartEN_2013_11.pdf

The **sample solution** is a model of an ideal answer to an exam task according to the person drawing up the examination. It has two functions. First, it serves as a reference for exam correction. Second, it is a reference for students who, after the examination, want to see how they could have approached an examination task successfully. This enables additional targeted and sustainable learning above and beyond the examination.

2.2 Correction and scoring

2.2.1 Rate the route or the result?

Both the result of a task and the route to that result can be scored. Scoring the result is usually simpler and more efficient. The students know what is expected of them with the task; in addition, unorthodox but effective solutions are rewarded appropriately. Scoring the route makes it possible to appropriately differentiate performance, especially if acquisition of competences has not yet been concluded or is incomplete. However, formulating clear scoring criteria is more demanding, and correction generally takes more time. This method also carries the risk that the person correcting the exam will not reward unfamiliar or disagreeable approaches appropriately.

2.2.2 Correcting examinations: Eliminating interference

Learning objectives describe what is being tested. All other factors are to be considered as interference, and their influence on exam scoring should be minimised. Examples of such interfering factors are the mood of the person correcting an examination, interpersonal differences in how people rate work, handwriting quality, and language skill. Please find a more detailed list of interfering factors in appendix 5.

The following six measures will help to limit the effect of interfering factors.

1. Examination papers should be scored **anonymously**. Paper examinations, for example, can be labelled with student ID and name in predetermined places which can be pasted over, covered, or folded away before correction takes place. Anonymisation is particularly easy in online examinations.
2. If correction is divided up among several people it should be **exam tasks** and not students' whole exams that are **distributed**. Scoring of each exam task by two persons independently of one another makes scoring more objective. In courses with many examiners tasks can be divided up among correctors according to expertise in the respective content.
3. **The order** in which student examinations are scored **should be varied** with each task.
4. **Taking conscious note** of potential **interfering factors** greatly limits their impact from the outset (see Appendix 5, 'Overview of interfering factors').
5. **Scoring charts and sample solutions** enable consistent and reproducible scoring of tasks according to uniform criteria and thus reduce the influence of interfering factors. The scoring chart and sample solutions should be discussed before examinations are corrected.
6. By using **reference corrections** the persons correcting the exam can 'gauge' their scoring of individual tasks. Corrections should be spot checked for consistency.

Interfering factors should never be retrospectively compensated. Doing so would comprise an arbitrary act and only puts the validity of the examination further in question, in that it merely adds a new interfering factor to the mix which has nothing to do with actual performance.

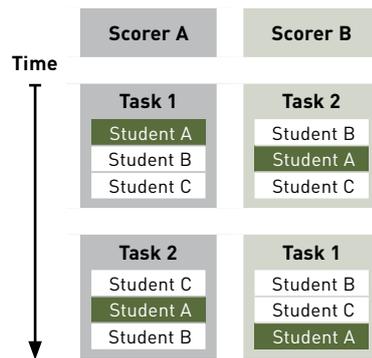


Figure 4 : Possible distribution of examination tasks and students among correcting persons

2.3 After the examination

2.3.1 Viewing the examination

The corrected examination contains differentiated information on the respective student's performance. Viewing it enables the student to use this information for further learning. Via critical and constructive student feedback it may also help to improve the quality of future examinations. Viewing the examination also has a legal function⁶. Retrospective corrections on the basis of exam viewing should be undertaken only rarely, with great forbearance. They are only indicated where scoring is clearly incorrect or in cases of severe discretionary misjudgements.

2.3.2 Reviewing the examination

The following steps are recommended in reviewing the examination:

- An **item analysis** helps to identify tasks which were too difficult or too simple, and pinpoints erroneously scored tasks for further checking (see 2.3.3).
- A **review of criteria** scrutinises the validity of the examination tasks and the scoring. It concludes by checking whether variations in student answers really correspond to differences in achievement of learning goals, and whether these differences can be reliably assessed using the respective scoring chart.
- A **comparison with earlier examinations** or with examinations from similar courses offers additional reference points for the correction process.
- A **group discussion** of the examination with students and/or Assistants can provide valuable feedback.
- Examinations are evaluated by students in the context of the **teaching evaluation** by ETH students. This evaluation provides important tips for the design of future examinations, in particular regarding the alignment of examinations with learning objectives and teaching, design of exam tasks, and fairness.
- Any necessary **changes in learning objectives** should be undertaken.
- The most pertinent findings regarding the examination's design, implementation, scoring and review should be compiled in an **examination report**.

⁶See Appendix 3: directive on 'Viewing and transfer of performance assessment records'

2.3.3 Identifying problematic questions: Item analysis

In an item analysis statistical values are computed for each task. They simplify the identification of tasks for nearer scrutiny.

- **Task difficulty** describes the proportion of students who have completed a task successfully. Too many too difficult/too simple tasks can impair the examination's reliability via floor or ceiling effects.
- **Discriminance** describes how performance in a particular task correlates with performance in the rest of the examination. Zero correlation signifies no relationship. The deeper a task's discriminance, the more probable are errors in the task's construction and/or scoring.

Good item analysis values are no guarantee of good tasks, however, and bad ones do not necessarily mean that the task is bad. The values provide hints, which must always be confirmed by a check of criteria.

2.3.4 Dealing with examination errors during correction

It can happen that in retrospect grave design errors are discovered or it becomes apparent that the examination was too difficult or too simple.

Examination tasks with **design errors** cannot simply be excluded from exam scoring after the fact. As a remedy of first resort, re-scoring on the basis of modified criteria will help. Here care must be taken, however, that the latter still accord with the scoring chart and criteria which were communicated to the students. If this is not possible, all students may be awarded the maximum number of points for the respective tasks, while ensuring that students who completed the task successfully either wholly or in part incur no disadvantages. If a task is re-scored the same good-practice rules apply as to the original scoring.

If substantial **doubt regarding the appropriateness of exam difficulty or of the grading scale** emerges through the grade distribution or for other reasons, the following actions are indicated:

1. The appropriateness and correctness of the examination should be checked in a review (see 2.3).
2. On the basis of the review, the scoring chart should be corrected and the tasks re-scored.
3. On the basis of the review, the grading scale should be redetermined.

In criterion-oriented examinations the grading scale is determined before the examination. Adjusted corrections, as described above, should be the exception. If they become the rule, the examination becomes de facto norm-oriented. If this cannot be prevented it is better to inform the students before the examination.

Chapter 3: Principles

3.1 Aligning examinations with teaching and learning activities

Learning is an active process in which new information is set in relation to existing knowledge. New competences develop from preceding ones via new experiences which spring from old. Examinations, teaching and learning activities should be aligned with each other in such a way that they provide optimal support to student learning.

3.1.1 All of a piece (Alignment)

Instructors should shape learning by specifying clear learning objectives, designing targeted activities and appropriately evaluating achievement of goals via the alignment in form and content of learning objectives, teaching-learning activities and examination tasks.

Alignment ensures that for the examination students will focus on mastering competences as recorded in the the learning objectives. The corresponding examination tasks will be perceived by the students as relevant, create positive incentives and thus foster motivation for learning. This in turn will cause students to study more intensively and with more endurance. Ultimately alignment also improves retention of what is learned by linking it with previous consolidated knowledge.

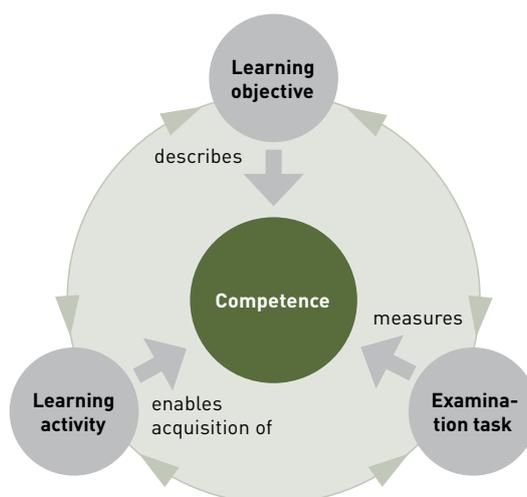


Figure 5: Well-attuned learning objectives, learning activities and examination tasks all relate to the same competences. This is called 'alignment'

3.1.2 Knew a lot, understood nothing? (Taxonomy of learning goals)

'Students are able to compute equilibrium concentrations of acid and base in aqueous solution.' In this example the **learning objective** consists of two components: a content domain (the 'learning matter', here equilibrium concentrations of acid and base) and a process related to the content ('compute equilibrium concentrations'). The achievement required now depends on two things: the retention of **knowledge** (facts; concepts; procedures or knowledge of knowledge: so-called metacognition) and the use of this knowledge (**cognitive process**), in this case the act of computation. The cognitive process describes the manner in which specialist content is processed: remembering, understanding, application, analysis, evaluation and synthesis. Simplified, this order of cognitive processes may be regarded as a type of 'processing depth'.

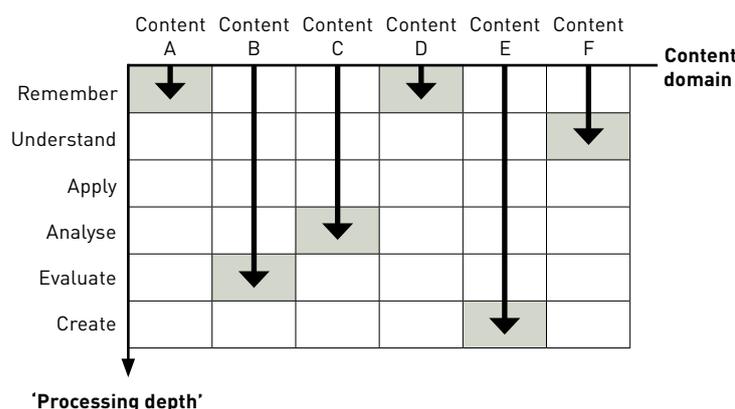


Figure 6: Learning objectives can be classified according to the expected 'processing depth'

Good examination tasks reflect learning objectives correctly not only in relation to their content domain but also in regard to the associated cognitive process. It makes little sense to test the learning objectives of a course where basic concepts are to be understood and applied via tasks which may be successfully addressed by pure memorisation. If, alternatively, the desire is to test whether students can transpose what is learned to new contexts, this must be made clear in the learning objectives.

3.2 Examinations are measurement instruments

Examinations are instruments which measure student competences. An examination surveys the achievement of learning objectives by first setting them out on a point scale and then on a grading scale of 1 to 6. To achieve the most meaningful results the aspects below should, as far as possible in teaching reality, be addressed.

3.2.1 When is an examination good? (Psychometric criteria)

An examination is good if it is capable of measuring student competences. This is never absolutely exact, because various measurement errors underlie every measurement. However, such errors can be kept small by complying with the criteria below.

1. The measurement must be valid, i. e. it must actually measure what it purports to measure (**validity**).
2. It must be precise and reproducible (**reliability**).
3. It must be independent of the examiner and the test circumstances (**objectivity**).

These criteria have a hierarchical relationship to one another. Objectivity is a necessary but on its own insufficient condition for reliability, and reliability in turn is a necessary but insufficient condition for validity. The most important criterion is always validity, because without it a measurement loses all value.

3.2.2 Examinations measure competences (Explanation of variance)

As opposed to simple physical size, such as length in metres, achievement of learning objectives is something 'latent' and not directly observable. It can only be measured indirectly, by inferring from completed examination tasks. Ultimately this always boils down to explained variance: the variance in examination performance represented in grades should correspond to actual variance in the achievement of learning objectives.

The measurement procedure can be divided into five basic steps:

- 1) The target competence is defined and made concrete in the form of **learning objectives**.
- 2) In **operationalisation** a suitable method for (indirect) observation of the latent property is developed. This corresponds to the drawing up of examination tasks and includes the scoring chart and the sample answers. The examination tasks are no longer directly related to the target competences but to the learning objectives deduced from them.
- 3) The examination is conducted, comprising the actual **data collection phase**.
- 4) In **quantification** the examinations are rated according to a scoring chart and mapped to a scale. So far this is a purely descriptive procedure.
- 5) In **standardisation** the points are set in relation to an external measure (norm) and displayed on a new scale. A differentiation is made between individual, social ('norm-referenced') and factual ('criteria-oriented') norm references. The grading scale displays performance normatively as 'very good', 'insufficient', etc.

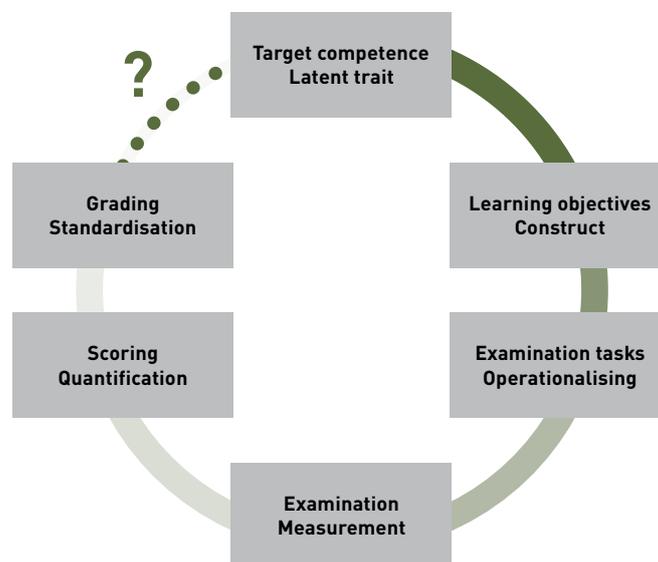


Figure 7: Competence measurement proceeds over several steps. Each of these steps introduces measurement error.

Here it must be kept in mind that each of these five steps introduces measurement errors. First, a part of the original variance is lost with each because it is no longer being tested: for example, in designing an examination, tasks cannot be included for all learning content or at all required cognitive levels. Second, error variance is introduced into measurement via interfering factors and imprecision of measurements: for example, factors such as the nervousness of students during examinations or their language skills affect the outcome. Variance lost in earlier steps remains lost in later ones, and error variance generally persists.

Error variance can, however, be minimised in two ways:

- **Interfering factors** can be identified and their influence minimised via appropriate counteractive measures. For example, language competence can be minimised as an interfering factor by always formulating examination questions in language which is as simple and clear as possible.
- **Unsystematic error variance**, or the 'fuzziness' of a measurement, can be minimised by introducing redundancy, or measurement repetitions. For this reason, where possible core competences or fundamental important facts should always be tested by more than one task.

3.2.3 Is a 6 twice as good as a 3? (Scale levels)

Measurements display properties meaningfully on a scale. Here the **scale level** determines how these measurements may be interpreted and which statistical operations may be performed on them. An **ordinal scale** depicts measurement values as hierarchically ordered rankings. Ordinal values may be compared with statements such as 'better/worse' and 'larger/smaller'. Examples of ordinal scales are surveys of satisfaction with a course unit, and the military hierarchy. Mode and median are defined as central values in ordinal scale data. An interval scale additionally allows meaningful comparisons of the intervals between values. Examples of interval scales are temperatures in centigrade, dates on the calendar and IQ. Defined as additional central values in interval scale data are averages (arithmetic means) and the operations of addition and subtraction. At ETH, passing an examination block and so gaining the right to continue one's studies usually depends on obtaining a pass grade measured according to a (weighted) average grade. Computation of the average (1) presupposes data on an **interval scale** and (2) requires that data contributing to the average also draw on a common scale in terms of what is measured. Measuring human competences via interval scales is both demanding and time-consuming, and cannot be reliably undertaken without the help of professional test developers. For this reason examination grades should in principle be regarded as ordinal scale entities. In addition, scales of different examinations may not a priori be used for others. From a psychometric point of view grade averages are therefore neither proper nor interpretable. This means that decisions on the basis of average grades are tainted by additional and hard-to-estimate error beyond the measurement mistakes of individual examinations. Nevertheless the average grade has become a heuristic in university practice which is accepted by both students and faculty, and it would be hard to imagine the Swiss educational landscape without it.

The more the individual examination grades based upon an average grade approach the level of an interval scale, and the stronger they reflect a common scale in terms of measurement content, the smaller the error introduced via formation of the average. To approximate interval scale a common metric must be determined for examination points such that one point always represents something comparable, independent of the task to which it is allocated. The relevance of a competence tested via an examination task, or the relative proportion of the learning objectives addressed might, for example, represent such a common metric. The resulting examination grade would then say something about the extent of relevance expressed by the competences shown in the examination. To ensure that various examination grades as far as possible reflect a common scale, it is essential that individual lecturers consult with one another and communicate on exam scoring.

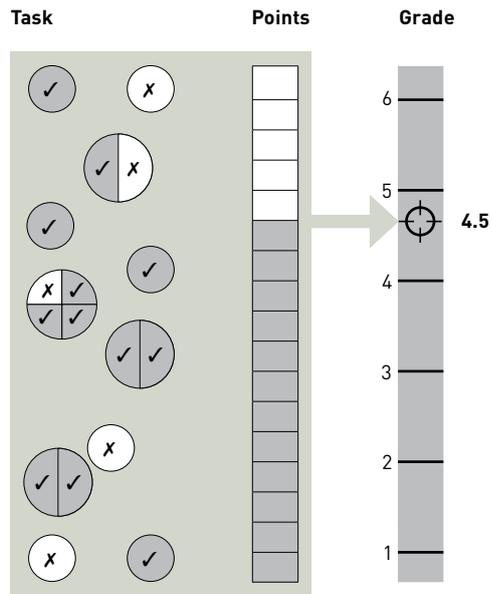


Figure 8: Tasks are corrected, scored and mapped to a grade. Because this is never perfect, particular care is required.

Appendix

Appendix 1: Correction and grading checklist

This checklist provides a brief overview of recommended grading procedures and will help with practical implementation.

Aspect	Central question
1. Preparation	
→ 2.1.7	Do I have a scoring chart?
→ 2.1.7	Do I have sample solutions?
→ 2.1.7	Has the correctness of the scoring chart and sample solutions been checked?
→ 2.1.2	Do I have earlier or similar examinations for comparison?
→ 2.1.1 – 2.1.3	Has the grading scale been determined?
→ Appendix 5	Do I have a sheet with an overview of the major interfering factors?
→ 2.2.2	Have examinations been distributed for correction according to tasks and not according to students?
→ 2.2.2	Has the order of student papers been varied during the scoring of each task?
→ 2.2.2	Were the scoring chart and sample answers discussed with my colleagues in advance?
→ Appendix 5	Has each task been scored by two persons?
→ 2.2.2	Have the examinations been anonymised?
→ 2.2.2	Have enough breaks etc. been planned into the correction process?
2. Correction/Scoring	
→ 2.1.7, 2.2.2	Do I have a scoring chart and sample solutions?
→ Appendix 5	Am I rested and fit enough?
→ 2.2.2	Have I reminded myself of the most pertinent interfering factors?
→ 2.2.2	Have I gauged my corrections by looking at reference answers?
→ 2.2.2	Have I scored as objectively as possible?

→ 2.2.2	Have I submitted scores without compensating for interfering factors in retrospect?
→ 2.1.7	Is the allocation of points, with explanations, documented in the examination?
→ Appendix 5	Has task scoring been checked periodically for consistency?
3. Viewing the examination	
→ 2.3.1	Has viewing of examination results and corrections been organised and conducted?
→ 2.3.1, 2.3.4	Have any necessary corrections been undertaken?
4. Reviewing the examination	
→ 2.3.3	Has an item analysis been conducted?
→ 2.3.2, 2.3.3	Have items with bad values been checked in terms of content?
→ 2.3.2	Have scores been checked?
→ 2.3.2	Has a discussion of the examination been conducted?
→ 2.3.2	Has the examination been evaluated in the context of the teaching evaluation?
5. Measures	
→ 2.3.4	Were there any doubts regarding the appropriateness of the exam level and grading scale?
→ 2.3.4	Did the need for new correction become apparent while reviewing the examination?
5.1 Re-scoring	If measures are required:
→ 2.3.4	Has the scoring chart been formally retained and only adjusted in content?
→ 2.3.4	Have the normal rules of good practice been retained in the new scoring procedure?
→ 2.3.4	Is the new scoring procedure consistent with the scoring criteria communicated to the students?
5.2 New grading scale	If re-scoring is impossible or was unsuccessful:
→ 2.3.4	Was the gauging of the new grading scale undertaken according to content-related criteria?
6. Quality management	
→ 2.3.2	Have any alterations been required and undertaken in the learning objectives?
→ 2.3.2	Have the most important findings of the examination review and the whole examination process been summarised in a report for next time?
→ 2.3.2, 2.3.4	If re-correction was necessary or the grading scale had to be adjusted, have preventive measures been taken to avoid a repetition?

Appendix 2: Further reading

Anderson, L.W., Krathwohl, D.R., Airasian, P.W. et al. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32: 347–364.

Biggs, J., Tang, C. (2011): *Teaching for Quality Learning at University*. Maidenhead/U.K.: Open University Press.

Hattie, J.A.C. (2002). What are the attributes of excellent teachers? In *Teachers make a difference: What is the research evidence?* (pp. 3-26). Wellington: New Zealand Council for Educational Research.

Eidgenössische Technische Hochschule Zürich (2013). Quality Criteria for Teaching: Section 'Degree programmes and courses'. Retrieved from: http://www.let.ethz.ch/docs/Qualitaetskriterien_LehreETH_EN.pdf

Krathwohl, D.R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41 (4), 212–218.

Krebs, R. (2004). Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Bern: Institut für Medizinische Lehre IML, Abteilung für Ausbildungs- und Examensforschung. Retrieved from: http://www.iml.unibe.ch/dienstleistung/assessment_pruefungen/pruefungsmethoden/wahlantwortfragen_mc/

Metzger, Ch., Nüesch, Ch. (2004): *Fair prüfen - Ein Qualitätsleitfaden für Prüfende an Hochschulen*. St. Gallen: Institut für Wirtschaftspädagogik, Universität St. Gallen.

Race, P., Brown, S., Smith, B. (2005). *500 Tips on Assessment*, 2nd edition. RoutledgeFalmer: New York, pp. 2–11, 21–22.

Schneider, M., Stern E. (2010). The cognitive perspective of learning: ten cornerstone findings. In H. Dumont, D. Istance & F. Benavides (eds.): *The Nature of Learning: Using Research to Inspire Practice* (S. 69–90). Paris: OECD.
doi: <http://dx.doi.org/10.1787/9789264086487-5-en>

Appendix 3: Further resources

'didactica' courses: Continuing education in university teaching at ETH and the University of Zurich:
<http://www.didactica.ethz.ch/>. A small number of courses are in English.

Guidance from LET: The LET guidance team is here to help you with all matters related to teaching:
beratung@let.ethz.ch

Guidelines for Lecturers found under:
<http://www.ethz.ch/faculty>

Directive on viewing and transfer of performance assessment records:
<http://www.rektorat.ethz.ch/directives>

Leistungskontrollenverordnung ETH Zürich [Ordinance on performance assessments at ETH Zurich]:
http://www.rechtssammlung.ethz.ch/pdf/322.021_leistungskontrollenverordnung_eth_zuerich.pdf (in German only)

ETH Zurich Ordinance on Performance Assessments: Implementation stipulations determined by the Rector:
<http://www.rektorat.ethz.ch/directives>

Appendix 4: Mathematical formulae

Definition of task difficulty:

Median number of points per task divided by the maximum achievable number of points per task

$$\text{Formula: } p = \frac{\bar{x}}{x_{max}}$$

\bar{x} = arithmetic mean (average) of the number of points achieved per task

x_{max} = maximum achievable number of points per task

Definition of discriminance:

Discriminance is the correlation of performance in a particular item with performance in the test as a whole without this item.

$$\text{Formula: } r_{i(t-i)} = \frac{\sigma(x_i, x_{t-i})}{\sigma(x_i)\sigma(x_{t-i})}$$

x_i : Values for item i

x_{t-i} : Values for the test as a whole without item i

Transposing to Excel:

The examples are based upon a table in which tasks are entered in columns A – O and lines 1–10 represent the students.

A1 – A10: Lines with students' points

A11: Field with the task's maximum achievable number of points

Q1 – Q10: Total sum of points of all tasks

Difficulty for task in column A:

AVERAGE(A1:A10)/A11

Discriminance for task in column A:

CORREL(A1:A10;Q1:Q10)

Appendix 5: Overview of interfering factors

Disruptive factor	Measure
<p>'External' characteristics of the examinee such as appearance, demeanor, participation and interest in the class, diligence, gender, etc.</p>	<p>The influence of these factors can be prevented by making examination papers anonymous to the scorers.</p>
<p>Aspects of the work handed in which have no direct bearing on actual learning goals, such as answer length, certainty, layout, and language skill.</p>	<p>Concrete, objective scoring criteria in the scoring chart restrict these influences.</p>
<p>Handwriting quality has a strong influence on task scoring.</p>	<p>This interfering factor can be minimised by asking questions which require relatively short answers and by conducting handwritten examinations without time pressure. Examinations on the computer eliminate the factor entirely.</p>
<p>Changes in the 'internal state' of the person correcting the examination generate inconsistent scoring: the tendency is to score more strictly in the early morning, if one is in a bad mood, and before midday if one is hungry. One is stricter with the first task to follow particularly good work. Conversely, scoring is more lenient before going home, if one is in a good mood, after lunch, and of an exam's last tasks if one is tired. Better scores are given to tasks which follow bad work.</p>	<p>These interfering factors can be lessened by taking regular breaks and by 'gauging' using reference scores. By scoring the same task student by student, and varying the order of students with every task arbitrarily or systematically, these interfering factors can be distributed more evenly among all students.</p>
<p>Different persons assess the same work differently. If correction is divided up among several people it should be exam tasks and not students' whole exams that are distributed.</p>	<p>The scoring chart and sample solution should be discussed before correction begins, to guarantee uniform correction. Scoring of each exam task by two persons independently of one another and at least spot checking of corrections for consistency help to make scoring more uniform and objective. In courses with many examiners, tasks can be divided up among correctors according to expertise or responsibility for the respective content.</p>

