Exam grading with the step-level model

The exam is written, the presentation given, the homework handed in. You've worked with the students the entire semester and now can read through the results of the exams to see how successful the learning process was. What you need for this is a reliable instrument to evaluate the exam performance: How do you recognize a 1.0 performance? Which results were barely enough for students to pass? The more concretely you can describe which qualitative competencies you expect your students to acquire, the more transparent, objective, and fair your evaluation will be.

Authors:
↗Michael Hagen, Birgit Szczyrba, Antonia Wunderlich (2015),

The process to measure students' performance in assessments is based on the model developed by Prof. Dr. Dr. Oliver Reis (↗oliver.reis@uni-paderborn.de).

We are happy to answer your questions or hear your suggestions regarding our information sheets!

Contact person:
Team Hochschuldidaktik
Zentrum für Lehrentwicklung (ZLE)
Team Academic Development
Coordinator Teaching A-Z
T: 0221-8275 3820
E: hochschuldidaktik@th-koeln.de

ZLE
Zentrum für Lehrentwicklung

Technology Arts Sciences
TH Köln

Exam evaluation for different taxonomic levels

For exams on the bottom of the taxonomy levels, the evaluation is quite easy: For example, exams requiring students to define technical terms are on taxonomy level 2 – understanding – and it is only relevant whether they a) correctly and b) completely recalled and reproduced the definitions. These types of exams can be evaluated on their identification of relevant content ("Which elements of the definition can be identified?") or a simple true-false format ("Were these elements correctly used?"). Alternatively, when students are asked to apply a formula, it's quite easy to see whether they correctly applied it or whether computational steps were left out. These evaluations are justified if you assign one point for every correct definition and every correctly computed result.

But grading systems should also be able to evaluate complex achievements and outcomes. A taxonomically sophisticated exam cannot always precisely measure tasks in the same way as can an exam with easier tasks on a lower taxonomy level, because the measurements "true," "false," or "perfect" admittedly play a role in your evaluation, but are not enough. To analyze a business strategy in marketing you need, for example, additional criteria such as "appropriateness," "creativity," "merit," or "data-based." In order to evaluate complex outcomes, it's not so easy to reduce your grading to points to a formulation that is connected with the requirements you have written in your ↗Learning Outcomes. Point systems designed for a lower taxonomy level steers your teaching in the wrong direction (see Info Sheet ↗Constructive Alignment) and are not transparent for students. The stampede to office hours ("Why did I get only a 3?") is already foreseeable.

How do you explain why one outcome is better than another? What did the students who "only got a 3" not do? You should be able to precisely describe which empirically observable and other comprehensible performance measures lead to a grade of "very good" or "acceptable." For this you need your ↗Learning Outcome to precisely define the competencies students should have acquired. You also need to ask students ↗Exam Tasks aligned with actions described in the Learning Outcomes to indicate which competencies they have acquired.

For dissertations and homework assignments, the exam keys used are often formulated with assessments such as "above average" or "consistently below expectations." While these assessments appear objective, they merely compare the performances of students in one particular group with one another. They lack understandable descriptions of student performance, and it's likely that for the same performances you assign different grades in different academic years. Finally, in a poorly performing group you must anyway give a 4 because there will always be students whose achievements are somewhat weaker than their fellow students. If in the next academic year the group performs worse, it could be that somebody in this group is outstanding who would have only been average in the first group – and then your exam results are inconsistent from one year to another.

An example

Suppose you offer a course in materials engineering whose ↗Learning Outcome lies on taxonomy level 3 (application):

WHAT Students can select materials for specific, technical applications in building services.
WITH WHAT In order to do this, you have to take into account technical material properties in the plumbing, heating, and air conditioning trades such aspects as valid standards, regulations, and laws for various building types (single-family buildings, multifamily buildings, hotels, hospitals, and industrial buildings). You apply aspects of cost-effectiveness, sustainability, and resource conservation.
WHAT FOR This makes possible the complex job-related planning and project-development tasks for aspects of building services and sustainability planning related to Green Building.

This also fits to the ↗exam task on taxonomy level 3: "Plan a cold and warm water installation for a hospital in Dinslaken. Hand in a strand schema and water analysis.

1. Which materials are allowed according to DIN 1988-200—drinking-water regulations for
   a. cold water?
   b. warm water?
2. Which accompanying material-selections regulations (German Association of Gas and Water Specialists, German Society of Engineers) need to be observed?
3. In your selection of materials, take into account the given results of the drinking-water analysis, hygienic aspects, and regional characteristics and special features.

For your chosen cold and warm water materials (at least 3 for each), provide a ranking for both the sustainability and cost-effectiveness. Substantiate your chosen material choices."

Identify your tools

For your step-level model, identify which tools the students need to apply in order to complete the task. Only by combining all the tools can the task be completed – which means that making a single mistake in the application of one aspect has an effect on all the other steps. The more taxonomically demanding the course is, the more the individual tools are interconnected, a characteristic of depth learning.

In our example pay attention to:
- regulations for the materials selection
- material characteristics such as corrosion resistance, material tolerance, connection technology, and temperature resistance
- building type
- requirements for warm-water installation
- water-quality standards and local requirements and characteristics
- sustainability (e.g., recycling rate, extraction of raw materials)

"In order for [your grading] instruments to offer descriptive indications [to students], the formulations of the scaled degrees of value should neither contain analytical judgements that conceal the actual criteria nor should they jump right into the evaluation – good, very good, fair, satisfactory, etc. Use formulations that you can point to as coming from your Learning Outcome, such as how the learning outcomes might look in practice. With these instruments you need to apply a keen observational filter to evaluate student outcomes using these criteria." [1]

Establish maximum, minimum, and average standards

In the next step, write down possible student-performance outcomes on three levels: maximum, minimum, and average standards. For the maximum standard, think about what a specific student outcome looks like, one completed in the ideal way, one you consider to be flawless. For the minimum standard, write clearly and plainly what students need to do to just barely pass. For the average standard, describe what outcome you expect for an average performance.

THE MAXIMUM STANDARD describes what an ideal and at the same time realistically achievable assignment outcome looks like. You assess how students who are still learning can meet and apply performance standards corresponding to the standards within their study program. The maximum standard corresponds to the grade "very good"; performance at this level will be practiced in the ↗Learning Rooms of your course and are transparent for all students. Sometimes in actual practice the grade "very good" is given only for outcomes that lie beyond the reach of your course (or students' abilities), a situation that is not transparent and unfair. In our example the maximum standard is described as follows:

The test taker…
- identifies the material choice using the relevant content from DIN 1988-200 and applies it correctly in combination with the permitted material
- takes into account in the task the accompanying guidelines from the German Association of Gas and Water Specialists and the Society of German Engineers
- can identify all materials for drinking water and differentiate the possible materials for cold- and warm-water applications
- plans out scenarios for additional differentiations of the material selection; how, for example, regional conditions affect the selection of the specified materials
- in their decision making, mention water quality and aspects of hygiene
- excludes certain identified materials based on their knowledge of material qualities, using criteria such as temperature resistance, corrosion resistance, and joining techniques
- outlines scenarios that further differentiate their choice of materials, such as regional characteristics
- consider and substantiate the advantages and disadvantages of their material selection from possible materials with reference to their sustainability and cost-effectiveness
- in all these steps displays a systematic approach to selection and justification

THE MINIMAL STANDARD describes – continuing with our example on taxonomy level 3 – how flawed an outcome in the area of "application" needs to be in order to just barely pass the exam. From the technical and scientific perspective, what is the minimum necessary? What student outcome represents the minimum necessary for further study and participation in the professional community without undermining the taxonomy level of the learning outcomes?

If the minimum standard can be achieved when students no longer "apply," but merely display "understanding" (taxonomy level 2), then the exam results are distorted: A grade of 4 indicates and translates to "the outcome is barely adequate in the sense of ↗Learning Outcomes" – it must also be an application of learning, otherwise the learning outcome is not achieved and student don't pass the exam.

The test taker…

- can to some extent explain and justify the choice of materials based on the content of DIN 1988-200
- choose materials based on practical experience but without detailed justification
- cannot differentiate the drinking-water materials section for cold/warm water applications
- takes into account neither water quality nor hygiene
- takes into account neither sustainability nor cost-effectiveness

THE AVERAGE STANDARD, that is, an average satisfactory performance, lies between the previously defined standards.

The test taker…

- can explain, justify, and link the choice of permitted materials based on the content of DIN 1988-200
- choose materials based on practical experience, although without detailed justification
- differentiate the drinking-water materials selection for cold/warm water applications
- takes into account water quality/hygiene or sustainability/economic concerns

If you still can't describe the range of student outcomes with these three step levels (maximum, minimum, average), provide additional intermediate levels. The differences between your differentiations should be ever more fine-grained. Develop five to six performance-step levels and always link them with taxonomy levels.

When you have "flawlessly" formulated your ↗Learning Outcomes, provide fitting ↗exam tasks. In order to make your evaluation criteria transparent for your "clients" – current students, your colleagues in the study program, and students in future semesters – you need structured guidelines for the development of your performance-level models with understandable definitions of tools and performance levels. Competence orientation requires a measurable and criteria-oriented evaluation. The work is worth it!

References

1 **Reis, O. (2014):** Nicht veröff. Weiterbildungsmaterial im Rahmen der Multiplikatorenweiterbildung "Kompetenzorientierte Prüfungskonstruktionen entwickeln" an der TH Köln 2014-2015.