Ethnography and Education, 2013
Vol. 8, No. 1, 89–104, http://dx.doi.org/10.1080/17457823.2013.766436

Practices of grading: an ethnographic study of educational assessment

Herbert Kalthoff*

Department of Sociology, University of Mainz, Mainz, Germany

The school as an institution assumes that students' grades are constituted by their assessments. This paper examines the background of this presupposition and provides a micro-analytical perspective of the grading practice of teachers in German High Schools (*Gymnasium*). This paper conceptualises the theoretical framework of the research in educational measurement in discussion. It is shown that the measured assessment of students and the teacher's observations are linked. When grading, teachers construct their own assessments. This process is depicted in this paper by two forms of observations: self-observation within the context of written examinations and third-party observation within the context of final oral exams.

Keywords: grading; human measurement; school; ethnography

1. Introduction

From a sociological perspective, the practice of educational assessment is interesting on two levels: on the one hand, the school conveys knowledge to students and on the other hand it allocates socially relevant knowledge about the students, making them comparable. Surprisingly, sociology and educational sciences 'know little about *how* teachers actually evaluate students' (Kingston 2001, 92; italics added). This paper is about this 'how': about the practices of teachers when grading, say, a written exam, and about the forms of knowledge the teacher uses. This paper tries to open the 'black box' of educational assessment from the perspective of sociology of knowledge and a practice theory approach.

Sociological research has underlined the connection between the grading that occurs in schools (i.e. academic success) and the conservation of the social structure of modern societies. This research was construed according to the social or economic reproduction theory and emphasised the significance of the school for the reproduction of class-specific differences (cf. Bowles and Gintis 1976; Giroux 1983; Cookson and Persell 1985). According to this model, the school imparts specific knowledge and capacities that are essential in the subsequent working environment. The theoreticians concerned with cultural reproduction have especially highlighted the significance of schools for the legitimisation of the dominant culture in all of its specific characteristics (cf. Bourdieu 1996). In a smaller empirical study, Bourdieu (1988) demonstrates how teachers are influenced by the cultural capital that they attribute to their students.[1]

Extensive research has been carried out on the practice of educational assessment in what I call the *old evaluation research*; several examinations revealed that the

*Email: herbert.kalthoff@uni-mainz.de

results of the educational assessment systems do not do justice to the quality criteria of objectivity, reliability and validity. A combination of experiments revealed that different teachers award different results when asked to evaluate the same exam script, regardless of the subject matter, and give different grades when re-evaluating the same exam.[2] A variety of researches also show that the judgement of teachers is influenced by personal information regarding the student, e.g. their social provenance, their general scholastic achievement or their gender (see, for example, Starch and Elliot 1912; Eells 1930; Day 1938; Finlayson 1951; Carter 1952). The main characteristic of this old research is that it removes the performance of exams from their local contexts. This *decontextualisation* is based on the assumption that a scholastic exam is a measurement procedure whereby those measuring are interchangeable instruments which (should) arrive at identical and thus valid results because they merely *copy* the results.

Instead, the *new evaluation research* suggests that the concept of validity has to be critically re-examined (e.g. Gipps 1999; Shepard 2000; McMillan 2003; Moss, Girard, and Haniford 2006); a focus is necessary which repositions teachers' assessments and grading within the local teaching and learning cultures (cf. Brookhart 2004). Verkuyten (2000) analyses the strategies of teachers to legitimise their grading and save their own face in the teachers meetings. Pryor and Torrance (2000) take on a social constructivist perspective; they show how assessment is a result of the interplay between teachers and students. Kain (1996) sees grading practices embedded within the teacher community and teachers' cooperation. Experienced teachers strengthen their position by introducing their norms of grading and former decisions to novice ('young') teachers. In this respect it is reported (cf. Georgiou 2008; Terhart 2011) that school achievement is treated quite differently. Experienced teachers refer to stable factors such as intellectual ability, gender and social background, whereas novice teachers express a more dynamic view of intellectual ability. Instead, Cizek, Rachor, and Fitzgerald (1996) show how grading practices depend on the setting (school, local area, cohort of students) and characteristics of the teacher (age, gender, experience, etc.). According to them, teachers tend to give good grades by including subjective and objective factors in their grading. Shay (2004) observed grading practices of students in an African university. She states that grading is a situated and interpretative process, influenced by the respective disciplines of the teachers, their experience and relationship to the students. These factors impair cooperation among teachers as they transform grading into a hidden and idiosyncratic procedure (see Cizek, Rachor, and Fitzgerald 1996).

In their longitudinal study, Filer and Pollard (2000) base the assessment of educational performance on the school and its wider social environment. That is to say, there is always more to the assessment process than the assessing person, the assessed person and the assessed knowledge; the 'more' is the social context, composed of, for example, the student's and teacher's identity, the social role expectations, the institutional settings and the generated classroom contexts, the expectations of families, and so forth. Taking into account this broad notion of 'context', the assessment process is conceived not as a neutral technique but as one influenced and constructed by a wide range of social factors and social actors.

In this paper, I attempt a *recontextualisation* of grading practices in schools and also try to illustrate how the social and symbolic order of grading is negotiated and established depending on the situation. The phenomenon of evaluation practices being embedded in school contexts is based on the observation that teachers reflect their own practices of knowledge transfer and testing while correcting and grading exams. Teachers are confronted with what students have learned in their classes, their ability to reproduce the subject matters, their ability to identify school knowledge with exam questions and their ability to represent knowledge in an adequate written or oral form. This paper also shows how teachers deal with these uncertainties, how they recognise the students' answers as correct or incorrect and how they defend themselves in front of their colleagues.

The empirical material reported in this paper was gathered through ethnographic *fieldwork focusing* on teaching and learning practices in German High Schools (*Gymnasium*) in the course of nine months; major parts of the *fieldwork* were carried out during 1992–1993 in two Jesuit Colleges located in the south-west and west of Germany (cf. Kalthoff 1997, 2006).[3] I was able to locate five teachers from various disciplines (biology, geography, German language and history) willing to give me access to their grading procedures and allowing me to participate in eight oral final exams (A-Levels; *Abitur*). I followed the students through their written examinations and their final oral examinations. I visited the teachers at their working desk, at home, or in school, where they performed the grading. After the oral examinations, I listened to them discussing the grades. These direct observations were supplemented by interviews with teachers and students. During the interviews, I asked how they (the teachers) organised the exam, how they prepared the questions and what kind of difficulties they faced when correcting the written exams or assessing oral examinations. I asked the students how they prepared for the exam, how they processed the examination, and about the importance of grades within their *peer group*s. The empirical data were analysed using the open coding procedure of Grounded Theory (cf. Emerson, Fretz, and Shaw 1995).

This article investigates the involvement of teachers in the examination process and discusses how they view their role within the evaluation process. The teachers were asked about mistakes for which they felt responsible and those for which they did not feel responsible, how they formulated their learning (or rather teaching) targets and how they reflected on the fact that they are confronted with the results of their own instructions (see below). In the following, I compare two cases: the teachers' assessment of written examinations as a procedure of implicit self-observation (1) and their collegial assessment of oral examinations as third-*party* observations (2). Although it is very important how students react to the grades, this paper does not examine this aspect (however, see Putwain 2011; Zaborowski, Meier, and Breidenstein 2011).

Excursus: the German school and grading system

After elementary school (grades 1–4 in most federal states) students enter the first phase of secondary school (grades 5–10) in one of three different types of school (*Hauptschule*, *Realschule*, *Gymnasium*), which are differentiated according to their emphasis on either practical or theoretical knowledge. The second phase of secondary school (grades 11–12 and 13) only takes place in the *Gymnasium* (High School) and leads to the higher education (college / university) entrance qualification (A-Levels; *Abitur*).

The German grading system is as follows: "very good", "good", "satisfactory", "sufficient", "deficient" and "insufficient". The grade is expressed in Arabic numerals: “1” stands for “very good”, “2” stands for “good”, “3” stands for “satisfactory” etc.; “5” and “6” are failing grades. Teachers can differentiate using “plus” and “minus”: a weak “very good” is a “1–“, a strong “very good” is a “1 +” etc. In the second phase of secondary school a system of grade points (0–15) is additionally used.

Grade	Insufficient	Deficient			Sufficient			Satisfactory			Good			Very Good		
	6	5–	5	5+	4–	4	4+	3–	3	3+	2–	2	2+	1–	1	1+
Points	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The system of grade points is used to calculate the final grades students receive for their *Abitur*; the grade points have to be differentiated from the points students receive for correct answers on single exam questions.

2. The exam as self-observation

One of the basic assumptions of school education is that students are continually and incessantly confronted with new knowledge (subject matter) and that, even though they are temporarily excluded, they can ‘follow’ the knowledge communicated. Students move within the framework of accumulating knowledge; their oral answers show ‘where the students are’ (to quote a teacher). The form in which the subject matter is dealt with can be denoted as *symptomatic knowledge testing*.

The organisation of school education has established moments of *systematic testing of knowledge*: the written exam. According to the general assumption, the written exam records the current situation of students regarding the subject matters dealt with in class in a certain period of time. The written exam thus allows classifying students on a scale according to their acquired knowledge. The timing of the classification is determined by the students’ knowledge: it is necessary to test their mastery of knowledge at a time when it can still be tested. That is to say that the school time consists of periods of acquiring knowledge and the subsequent evaluation of this acquisition by way of regular oral and written exams.

2.1. The Assessment System

During the exam, the students produce a text which is given to the teacher. The teacher then assesses the quality of the text. The process of grading the text is not just one of ticking off right or wrong answers, i.e. observation of another person’s performance. On the contrary, the grading is problematic and teachers are ambivalent about the process because it might jeopardise the conception that the teacher knows his or her students and that they know their subject matter well. The uncertainty of the results, which can be a disaster for the teacher, is the prerequisite for alternatives of ascription and for the phenomenon of self-observation.

This problem of (not) appearing to know their students and the subject matter is met by teachers in several ways. They tend to procrastinate regarding the grading process and usually correct them at home, rather than at their desks at work. Thus, they remove their work from other people’s sight (cf. Shay 2004). Teachers enter an exam situation themselves when evaluating the students’ texts. At the same time, teachers may have doubts as to whether they are reading representations of students’ knowledge: a correct answer from a student does not necessarily mean that the student ‘has understood it’. This is a basic doubt about the function attributed to exams, namely that of measuring a performance. In these cases, teachers are not reflecting upon their own teaching competence but rather upon the exam’s representative function. As Peter, one teacher, observed in his class (Geography, 11th grade):

> When they draw the circulation system of land and sea breezes they are actually showing that they know their stuff – *actually*. Maybe they only memorised it though, you never know, but they didn’t describe it explicitly. I can memorise the circulation system without understanding it, right? A lot of students do that. But I can’t test that.

Interviewed teachers were aware of the impossibility of testing whether the students understood or memorised the knowledge documented in their answers. And they only had one possibility of reacting: They had to award the student the point (score).

In practice, assessment systems can be differentiated according to how teachers organise their grading. Teachers using the ‘criteria- and learning target-oriented’ procedures were observed; they constructed model answers before they began grading with which they compared the students’ answers. These model answers combined the elements of knowledge that teachers hid in their questions with an evaluation template that is expressed in numbers (‘points’). The amount allocated for a particular component of the exam indicated the degree of difficulty and significance of the solution. The model answers transform the embedded knowledge into relative numbers in order to calculate overall scores (which are treated as directly corresponding to a student’s overall knowledge) economically; the written knowledge documented by the students became balance sheet material. Grading using model answers is basically a comparison between the balance due (knowledge) and the actual balance in the students’ texts. The model answers are a reflexive instrument; they demonstrate a required level of knowledge that is independent of the students’ knowledge. This calculatory grid of the distribution of scores needs to be distinguished from a ‘group-oriented’ procedure. For example, at the beginning of a grading session, a teacher – Paul, teaching German (7th grade) – asked himself what he expected to read in the exam responses. He took notes without assigning points to the answers. In other words, a vague list of expectations served as a guideline throughout the grading process.

Other teachers ignored this visualisation of their expectations completely and only concentrated on the information provided by the students in their written exam. These were often cases in which the teacher’s grading prospect was defined by the students’ performance. The performance of the students already known to be ‘better’ was then considered the *benchmark* performance. Some of the teachers systematised the procedure by making ordered piles of written exams before they started grading: top, middle and lowest level of (expected) achievement.

The model answers stand for a fixed norm based on the questions asked; they *preclude* the grading process whereas the ordering of the exam scripts is a type of measurement that takes place *after* the exams of the supposedly better students have been graded. There are therefore different patterns of selection methods; first, they differ regarding what they refer to, to right answers or to expected outcomes.

Secondly, they differ regarding the question as to when the student performance has to be integrated into the grading, at the end or at the beginning of the procedure.

2.2. The procedure of distribution

Before they start the first run-through, the teachers take a look at the exams script: the text quantity was determined or estimated and a few exam papers were leafed through. They noted orthographic, grammatical and content-related mistakes using red ink or a pencil. Points were noted in the margins next to the tasks in pencil, or on a note pad, or in a cross-classified table; notes were taken for the final judgement on scrap paper, e.g. 'elaboration of diagram missing'. There were different abbreviations used in the margins, for example, 'Gr' (grammatical error), 'FE' (factual error), 'FD' (factual deficiency), 'L' (logical mistake), '+' (good, correct), '?' (unclear), as well as carets, curlicues under words and notes (e.g. 'legibility', 'argumentation?!'). The teachers who worked with model answers continually glanced from the students' texts to the model answers as if they wanted to check the oscillation of the gauge.

During the often very long process of reading and grading the exam scripts, some of the teachers regarded the text from a (fictional) dialogical angle, as if the student were in the room. For example, one muttered: 'OK, dear John, what am I going to do with you?' The texts evoked sentiments such as happiness about the student's successful performance in the exam, surprise about an unexpected positive result and disappointment because a student was unsuccessful in the exam. Teachers suffered during the process. They sighed and, sometimes with a distant look in their eyes, they often showed a lack of understanding regarding that which they were reading. One said: 'What is she writing about?' (Lynn, Biology, 12th grade). Another complained: 'He completely misunderstood the task' (Peter, Geography, 11th grade).

When teacher finished reading a successful exam text and moved on to a less successful one (or vice versa), they tended to change their reading and correcting process. Correcting a good exam text was like reading in a concentrated way – either the pencil or the forefinger marked the line the teacher was reading. However, reading and correcting with red ink tended to go together when a bad exam text was being graded. The teachers seemed to have a libidinous relationship with good exam texts, and they treated them well, whereas bad exam texts were always treated rigorously. The teachers' sentiments indicated that they were confronted with their success (or failure) of transmitting knowledge to students. One can say that they were working on the ascription; the effect tended to be a rejection or approval of the particular student, with which the teacher expressed his or her identification or depreciation. In negative cases, the teachers did not seem to be evaluating, but rather devaluing. Thus, the teacher tended to decide whether the unsatisfactory performance was their problem or not: Who was responsible for the success or failure? They answered this question using their knowledge of the subject matter, the exam and the class itself.

In some respects the teachers looked for reasons for bad student results. For example, they might infer that a poor mark might be the result of their (the teacher's) bad selection of a textbook, or their insufficient explanation in class, or an unclear formulation of an exam question or that their expectations were simply too high. The student's insufficient preparation was followed by unambiguous finger-pointing. In cases like these, teachers can punish all of the students, or only particular students, with bad grades. Teachers also know that there are certain topics of subject areas (e.g. particular topics in chemistry class) in which exam results tend to be poor. In these cases, the bad results are attributed to the topic itself; when a new topic is taught, the poor results are 'compensated', so remarked a teacher. Self-observation also referred to the process of reading the exams or tasks repeatedly.

> You look at it again and again. Maybe he did get something right. And then he writes something *completely incorrect* or *pretty* incorrect, and you accept it anyway. (Peter; Geography Class, 11th grade)

> OK, well *there* is the context. I know that's what she means, but it's still in the incorrect technical terminology. But I can give her half a point. I can do that. (Lynn; Biology Class, 12th grade)

The intense (quasi hermeneutic) reading was a resource for further interpretation; the teachers used in-depth reading to find correct elements of the answer in the students' texts. Teachers make something that seems impossible to appraise appraisable; they shift the limits of 'still appraisable' and 'no longer appraisable'. Whether they can still 'bring out the best' in a student's texts depends on whether they find an appraisable element that can be expressed in scores. Teachers tended to decide in favour of their students. They often accepted an incorrect answer as correct and showed benevolence towards the exam text and the student. This benevolence was also obvious when the grades were adjusted to match the actual performance of the students between the first and second run-throughs.

At the end of the first run-through, the total number of points is jotted down in a cross-classified table and a first commentary including an approximate grade area was delineated. In the cross-classified table, a grid of students was established, sorted by elements of the answers, tasks and a total number of points achieved. A preliminary justification for a judgement call was made, such as: 'Martin will get a "very good". Even if his exam isn't actually in the area of "very good", he'll get a "one". Then it'll just be a weak "one" (a "one minus")' (Lynn; Biology Class, 12th grade). The second run-through has a monitoring function; teachers check to make sure that they have not overlooked elements of the answers and whether they have noted the right number of points. One teacher explained that this process was most intense for the exams of students 'in jeopardy' of failing because in this case the grading had to be performed very carefully. The total number of points were ascertained and compared to the result of the first run-through; a final comment was composed or written directly underneath the student's text. The repeated run-through reassured the teacher that no mistakes had been made and helped him or her monitor the distribution of the grades. Thus, the assessment was consolidated.

2.3. Working on the average grade

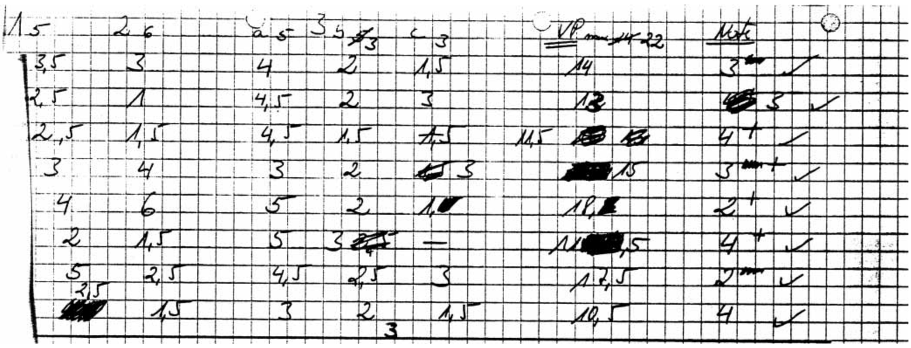
The fine adjustment of the single student's grade differed from the work that was conducted in respect to the average grade. If the teacher realised during the grading process that the exam results were 'catastrophic' (teacher) or that a certain task was not answered correctly by the majority of the students, teachers redefined their standards. This is the point at which, not only the students, but also the teachers themselves were under pressure. One teacher explained while grading a class test:

> And about the weighting of the exam, I can already say . . . If the first part was solved much better by the students than the second part, and if I'm in *danger* of having to give more than half of them a five (failing grade), then I will take the liberty of changing my weighting, my criteria. (Robert; History, 10th grade)

The need for adjustments developed if the result no longer fitted. Independent of the model answers prepared, good students functioned as an alarm signal for which the teachers listened. If the first run-through gave the teacher the impression that a bad result was likely, the teachers shifted their weighting and thus the average grade of the exam. Thus, they not only changed the distribution of the students' grades, but also shifted all of the students into higher grade areas.

> (1) For his educational assessment, the teacher (Peter; Geography Class, 11th grade) has chosen to follow the 'criteria-oriented' procedure. He compiles model answers, dividing 24 points among five exam questions. After having finished the first run-through of eight exams he realises that the answers to one particular question do not fulfil his expectations after recording the complete number of points, which he jots down in a table. 'After the first exams I realised that they can't reach all five points. Nobody achieved them'. The exam of a 'very good student' (teacher) is an important impulse to readjust the model answers in keeping with the actual answers given. The student was only able to give an unsatisfactory answer. When the teacher, who gets his initial bearings from the subject matter and his learning goals, realises that the results of the exam are going downhill, he reorients himself. He then goes back to the historically old selection method that starts off with the performance of the best students. This teacher justifies himself: 'I did it (changed the weighting), because the average of the exam was no longer adequate in terms of my expectations'. The teacher is insecure about the reasons, but takes over a large part of the responsibility: 'This shows that I either misjudged it or I just didn't explain it well enough'. (Fieldnote, May 1993).

In his model answers, the teacher had itemised the distribution of points with reference to the individual elements of the answer. While reworking his distribution of points, he crossed out the two elements that the students had not mentioned (see Figure 1).



1 5	2 6	a 5	3 5 ~~4~~ 3	c 3		VP ~~14~~ 22	Nte	
3,5	3	4	2	1,5		14	3 [redacted]	✓
2,5	1	4,5	2	3		13	[redacted] 3	✓
2,5	1,5	4,5	1,5	~~1,5~~	11,5	[redacted] [redacted]	4+	✓
3	4	3	2	[redacted] 3		[redacted] 15	3 [redacted] +	✓
4	6	5	2	1,[redacted]		18,[redacted]	2+	✓
2	1,5	5	3 ~~2,5~~	—		11[redacted],5	4+	✓
5	2,5	4,5	2,5	3		17,5	2 [redacted]	✓
2,5 [redacted]	1,5	3	2	1,5		10,5	4	✓
			3					

Figure 1. Work on the average.

Consequently, the students were shifted into higher grade levels. The possibility for readjustment is a method of production for 'adequate' results. The teacher accommodated the students' overall performance during the second run-through.

> (2) In the second example, another teacher (Robert; History Class, 10th grade) is using the 'group-oriented' procedure. First, he determines the text quantity (eleven students produced 34 pages of text), and then he looks at the exam topic once more. The teacher takes notes on topics concealed within the questions. There are two exam questions which the teacher would like to weight as follows: one-third (first question) and two-thirds (second question). After he sorted the exams ('manageable morsels') into grade areas – highest grade area, four exams; middle grade area: three exams; lower grade area, four exams – he starts the grading process. He begins with the pile of the highest grade area in order to ascertain the criteria for his distribution. These texts are meant to function as points of orientation for the distribution of other students' performances. He explains: 'I like to get my bearings from the good students, it's basically feedback'. But the exam of 'best student' (teacher) in the class is not as good as expected. The teacher assesses the first part of the exam as 'three minus' or 'four plus'. However, in the second run-through, the teacher gives the student a 'three' on the first part of the exam so that the student will have an adequate grade overall – 'I can't give him a 'four'' (teacher).[4] The second part 'has to be a 'two'', so that the teacher can justify an overall grade of 'two minus' or 'three plus'. After having read the four texts in the pile of the highest grade area, that is, the exams that were preselected in comparison to the other exams, the teacher gives the student a 'three plus'. After the first run-through, the teacher has obtained a good impression of the texts. He shifts his prior weighting of one-third to two-thirds to an equal value of both parts. That is, the part of the exam in which the students in the top grade areas performed better is given more weight. (Fieldnote, June 1992)

This example shows that the actual performance of the students is an indicator, but other indicators are the grades given in preceding exams and the prior performance of the student. As a consequence, the teacher stabilises the hierarchy of the students by using prior results as a current distribution key. The result of this procedure is that the students have reduced chances to move around on the grading scale.

3. The exam as third-party observation

When teachers are grading written exams, they are alone. The constellation consists of one teacher on one side and a defined number of students on the other. The fact that teachers always reflected upon their own performance when grading the students' exams – as postulated above – becomes obvious when the school reverses this constellation. This was, for instance, the case during the final exams. In the written part of the exam, a second and sometimes even a third teacher as corrector was involved (consecutively). The second and third teachers knew how the exam was evaluated by the first or second correctors.[5] The oral final exam was different; the grades were discussed in a group of three or four teachers with different functions: *chairperson*, examiner, assessor and minute taker. The *chairperson* was responsible for the overall procedure, the examiner was the teacher of the student, the assessor observed the teacher and is therefore generally a member of the same discipline (say physics) and the minute taker was responsible for an accurate documentation and wording of the oral examination. In the following, the analysis concentrates on the oral final exams which were part of the German *Abitur* (A-Levels) to describe the practice of the collegial exam as third-*party* observation.

3.1. The exam conversation

After a short briefing, the examinees begin with his/her presentation. The students' presentations in the exams observed took about 2–13 minutes. The examiner only interrupts if the examinee stalls or is 'on the wrong track'; in such moments, they give small hints so that the examinee can continue his presentation. As examiners, teachers do not tend to interrupt until the examinee marks the end of his presentation. The first part of the oral final exam inverses the regular teaching situation as a ritual of reversal; the teacher no longer has the right to speak first, the examinee does. However, for the examinee, the right to speak also means the obligation to speak. In class, he only has to listen; now he must speak: it's his *turn*.

When the examinee finished his presentation, an exchange – question and answer – between examiner and examinee followed, leaving the other teachers partially anonymous. The examiner adjusted his subsequent inquiries depending on the quality of the examinee's presentation. Some teachers were observed asking a sequence of questions without referring to or commenting on the examinee's answers. In this case, the examinee provided catchwords which led to further inquiry on the part of the examiner – or to a new topic.

> ER: . . . How does the underdevelopment come about?
>
> EM: There are different possibilities for it to come about. For example, the country can be dependent on products and prices, such as crude oil. And when the crisis came around and the prices for crude oil went down, the revenue of the countries went down as well.
>
> ER: Which role does the export of resources play?
>
> EM: For one, there is the possibility that a country has no resources, for example if the agricultural structure is bad; many products can't be cultivated. And the conditions in former times made a development almost impossible for those countries. . . . (ER = Examiner, EM = Examinee)[6]

A specific characteristic of this form of oral examination is that the examining teacher merely acted as a stimulator of answers and did not evaluate the examinee's answers through his further inquiry, comments or suggestions.

Other teachers organised the exam conversation in the form of a *cluing*-process (cf. McHoul 1985, 1990). This describes a procedure in which teachers give hints, catchwords or make further inquiries, thus giving the examinees keys to the solution of the problem (in this case, the exam questions). The correction of the examinee, that is the production of an adequate answer, is structured through a repeated cycle of the teacher's question – student's answer – teacher's question. Teachers using this procedure seemed to want to 'bring out the best' in the students, especially during the examination of a weaker student. However, the inquiries and hints have a different meaning depending on the knowledge the students showed during the examination; comments sometimes indicated that the examinee's answers were inadequate and that there was a need for a more correct answer. In other cases, comments were used to test the real knowledge level of 'good students'.

Both forms of exam conversation were identifiable as to whether and how the teacher (examiner) gave direct feedback to the examinee (or not). This depended on the correctness of the answers. Examinees able to answer the examiner's questions adequately are thus – in the sense of the examination ceremony – examinable, even *well* examinable; yet they were deprived of an enciphered evaluation of their performance. This also meant that the weaker examinees receive more feedback. However, the feedback was counted against them during the determination of the grade, because the feedback can be interpreted as 'help'. If the feedback was postponed, the examinee had a better final grade.

3.2. The making of the grade

Before the grade is determined, the examining teacher gives a brief summary of the course of the exam, i.e. the examinee's performance and the examiner's questions. The summary is followed by the other teachers' comments and additional observations, confirmation or corrections.

During these sometimes quite long discussions, the teachers I observed in the eight oral exams elicit what just passed. It seems as if these teachers needed to establish a collective certainty regarding 'what was there' or 'what was missing'. They present themselves as having searched ('there was nothing'), compared ('that was deeper (the knowledge) in the other exam'), worked hard ('I really had to drill into him') or as having judged aesthetically ('at times it was a little bit hard-bitten'). They assert the examinee's competence: he 'described', 'named', 'wrote down', 'implemented', 'saw' or 'discerned' – these are all faculties teachers are able to detect. Additionally, they reappraise the level of independence: to what extent did the examinee need the teacher's 'help' to present what he/she did ('in the second part she wasn't as good, she needed some help from you') or that of other media (e.g. a formulary in mathematics). They review the language ('in German we could understand well'), the self-assessment of the examinee ('at the end he started shaking, that's when he realised he might fail') and they read up on characteristics of the examinee's scholastic career ('kind of bumbled through'). Part of the complete impression also includes information regarding the grade the examinee 'needs', meaning whether he is in danger of failing the final exam and graduation (see example below).

However, not only the examinee's performance was in question but the examiner's work was debated as well. If teachers discussed their colleagues' course of action they demonstrated that they had intimate knowledge regarding the constructedness of the student's performance. The level of difficulty of the exam task to be prepared as well as the examiner's questions was considered challenging ('questions were pretty difficult') or easy ('*none* of it was particularly difficult'). The examiners risked having to justify their course of action either due to the quality of the examination question given or because of the answers given by a weak student. In these cases, other teachers started wondering 'how did the student ever manage to pass the course'.

In the following example, teachers discussed the grades of three students after their oral exam in physics; they worked chronologically and at this point were discussing the performance of the second examinee.

Example: Determination of the grade[7]

1 C: So, where would you place him?
2 E: I would say a three plus.
3 A: Three plus?
4 E: When I think of what he offered, then a three is too harsh.
5 M: I would give him half a point more.
6 C: Two minus?
7 M: Exactly.
8 E: We can't incorporate the things I didn't ask about.
9 A: None of it was particularly difficult.
10 E: I admit that, I stayed at the surface.
11 A: I would give him a plain three.
12 E: A plain three?
13 A: This is a major course.
14 E: I know, but he was very insecure. If you keep in mind that he had
15 physics as his major course, we can expect him to remember
16 some of it. And he did.
17 C: Should we call it a three?
18 A: If you can live with that decision. It's just my opinion. I'm just
19 speaking for myself.
20 M: So what does he need?
21 E: In both of his major courses he has a four
22 minus in one and a five plus in the other.
23 A: That would give him twenty-one or twenty-two points if we
24 give him a three.
25 C: So should we give him a plain three?
26 A: Yes.
27 E: Wait a second. I had suggested a three plus.
28 C: That would be eight points, with the four minus that would be
29 twenty-one points.
30 C: He has to have one, and he has that now. In physics he has
31 satisfactory (three) overall.
32 E: Huh? But we didn't
33 C: Here in the oral exam, this is where he has the three
34 E: All right…

The example demonstrates a controversial negotiation of a grade in the case of a student that might still fail and not be able to graduate. The first part (lines 1–17), up until the *chairperson*'s first attempt at mediation, was an exchange of grade suggestions (lines 2, 5, 11) and questioning each other (lines 3, 6, 12), followed by reasons for the grade suggested (lines 4, 7, 8, 13–16). The difference between the observation of the examiner and the assessor becomes obvious. The assessor had an abstract demand ('a major course', line 13), while the examiner, supported by the minute taker, argued on the basis of a mixture of formal criteria (line 8) and social dimensions ('very insecure', line 14). The examiner and the minute taker tried to link the assessment to the knowledge context established by the task and questions; it is not possible to evaluate something that was not required of the examinee during the exam. Yet the objection of the assessor, who criticised the level of difficulty as being too low, could not be rebutted by the arguments brought forth by the examiner and the minute taker.

The *chairperson*'s first attempt to bring about a consensus on the lower grade (line 17) failed. Since the line-up of grade suggestions did not lead to a result, more information on what the examinee needed is collected. The minute taker's question ('So what does he need?', line 20), followed by the examiner's elucidation, was succeeded by the assessor's 'that would give him…' –answer (lines 23–24). At this point, the *chairperson* again attempted to mediate, based on the suggestion made by the assessor (line 25). While the assessor endorsed this attempt (line 26), the examiner again tried to channel the discussion towards a better grade (line 27). The *chairperson* however ignored this intervention and unflinchingly continued calculating the number of points (lines 28–29).

In the final part of the discussion (lines 30–34), the *chairperson* was able to assert his first attempt at mediation, being the assessor's suggestion. His drive towards a decision ('he has that now', line 30) was followed by the examiner's repeated effort to make his suggestion heard and avoid the lower grade. In the interrupted sentence, he was going to complain that the teacher had not voted on the grade (line 32). The *chairperson* insisted on his suggestion (line 33), whereupon the examiner relinquished his opinion ('All right', line 34).

In this example, two pairs were up against each other: *chairperson* and assessor vs examiner and minute taker. Both pairs worked together while dividing the labour. One of the participants of the pair was responsible for the argumentation regarding the contents; the other was responsible for attempts at mediation or informative questions. The division of labour was visualised by the assessor's demonstrative withdrawal from the part of the conversation during which a decision was made; after having relativised his judgement, he lefts the enforcement of his suggestion up to the *chairperson*. This discussion elucidates the fact that the examiner identified with the examinee due to his obvious orientation towards the level of achievement of the examinee and due to his introduction of the social dimension.

4. Conclusion

This article has dealt with the practice of educational assessment by teachers and the constitution of scholastic judgement. Two procedures, individual evaluation and collective evaluation, were distinguished. By defining individual texts as being good or bad on a point or grading scale, teachers place one student's text in relation to another's. Teachers determine the gaps between students; they establish an order by placing the students at different levels. Subsequently, the students obtain an allotted position within the class and in the classification. The exam can only fulfil its differentiating function if the grades differ. Teachers consider it unthinkable to award only one or two different grades; their perception is that there are good students and those less than good. That also means that the ascertainment of the students' performance is dominated by the function of distributing the students' marks on a scale. The students are supposed to be *distributed* across the length of the scale; this is the purpose of the distributive work. For students, this means that there is no possibility of all of them being 'very good'. Equality is not the goal, but rather variance, which calls for social validity. Variance leads to the materialisation of the socially selective function of the school.

The path of the exam is similar to a multilevel reduction process. The intensity of the drafting process, the atmosphere in class, the reflexivity of the teacher regarding his or her own teachings as well as the requirements specified in the curriculum define the context of the exam. The path has several openings which facilitate adjustment to the norm, for which the teachers feel responsible.

The social effect of educational assessment practices permits further deliberations going beyond the empirical findings. These are linked to a second motif aimed at producing an order in the sense of an objectification. For assessment practices, this means that teachers do not only evaluate individual students' performances, but also organise a distinct distribution on a scale. Through institutionalised processes of knowledge testing, teachers contribute to the identity formation of students and make them distinguishable and recognisable. Ironically, due to its social claim to validity, educational assessment is sealed off and considered a fact that ignores the context in which it was generated.

Furthermore, it has been shown how teachers are involved in the process of assessment as generators of scholastic judgement; by reflecting upon the students and upon themselves, they always evaluate their own performance as well as that of the students. The self-construction of those performing the evaluation can be defined as a constructive mechanism of the scholastic process of differentiation. At the same time, the practice of assessment is a social practice. It does not follow a private logic although it is in most of the cases generated in closed rooms and thus systematically removed from public access. Teachers performing the practice by which they evaluate students and their own achievements place emphasis on the fact that the outcome of the process (the grade) can be represented as rational and justified by the 'symbolic power' (Bourdieu) of the teacher. The findings of the paper lead to a shift of perspective from the normative notion of performance measurement of students to the concept of a system of third-*party* and self-observation.

Notes

1. The concept of cultural capital has been critically discussed in the sociology of education. These studies presented ambiguous and puzzling results. On the one hand they show 'that cultural capital is positively related to high school grades' (DiMaggio 1982, 194) and on the other that the concept has to be respecified or redefined (De Graf, De Graf, and Kraaykamp 2000) according to social class, gender and national school systems (e.g. De Graf et al. 2000; Kingston 2001).
2. Just one example: Weiss (1965, 148–151) states that 153 Austrian teachers evaluated the same math test with grades ranging from 'very good' (7%), 'good' (41%), 'satisfactory' (42%), 'sufficient' (9%) to 'deficient' (1%). Hartog and Rhodes (1936, 9) describe findings wherein 28 teachers graded the same Latin exam in very different ways.
3. The overall aim of the research project consisted in analysing elite education in Germany; this will not be explored in this article.
4. Not being able to give a 'good student' a bad grade is an especially extreme example for the thesis postulated here, namely that the teacher's evaluation constitutes the student's performance.
5. Of course, this can lead to surprising and disagreeable results for teachers. However, this topic will not be explored in more detail here.
6. The excerpts documented in this paper were reconstructed on the basis of notes I took while following these discussions.
7. The transcription symbols are as follows: E = Examiner; C = Chairperson; A = Assessor; M = Minute taker.

References

Bourdieu, P. 1988. *The Categories of Professional Judgment. Postscript in* Homo Academicus. Cambridge: Polity Press [originally published 1975].

Bourdieu, P. 1996. *The State Nobility: Elite Schools in the Field of Power*. Stanford: Stanford University Press [originally published 1989].

Bowles, S., and H. Gintis. 1976. *Schooling in Capitalist America*. New York: Basic.

Brookhart, S. M. 2004. "Classroom Assessment: Tensions and Intersections in Theory and Practice." *Teachers College Record* 106 (3): 429–458. doi:10.1111/j.1467-9620.2004.00346.x.

Carter, R. S. 1952. "How Invalid Are Marks Assigned by Teachers." *Journal of Educational Psychology* 43 (4): 218–228. doi:10.1037/h0061688.

Cizek, G. J., R. E. Rachor, and S. F. Fitzgerald. 1996. "Teachers′ Assessment Practices: Preparation, Isolation, and the Kitchen Sink." *Educational Assessment* 3 (2): 159–179. doi:10.1207/s15326977ea0302_3.

Cookson, P. W., and C. H. Persell. 1985. *Preparing for Power. America's Elite Boarding Schools*. New York: Basic Books.

Day, L. C. 1938. "Boys and Girls and Honor Ranks." *The School Review* 46 (4): 288–299. doi:10.1086/440217.

De Graf, N., P. De Graf, and G. Kraaykamp. 2000. "Parental Cultural Capital and Educational Attainment in the Netherlands. A Refinement of the Cultural Capital Perspective." *Sociology of Education* 73 (2): 92–111. doi:10.2307/2673239.

DiMaggio, P. 1982. "Cultural Capital and School Success. The Impact of Status Culture Participation on the Grades of U.S. High School Students." *American Sociological Review* 47 (2): 189–201. doi:10.2307/2094962.

Eells, W. C. 1930. "Reliability of Repeated Grading of Essay Type Examination." *Journal of Educational Psychology* 21 (1): 48–52. doi:10.1037/h0071103.

Emerson, R., R. I. Fretz, and L. L. Shaw. 1995. *Writing Ethnographic Fieldnotes*. Chicago: University of Chicago Press.

Filer, A., and A. Pollard. 2000. *The Social World of Pupil Assessment. Processes and Contexts of Primary Schooling*. New York: Continuum.

Finlayson, D. S. 1951. "The Reliability of the Marking of Essays." *British Journal of Educational Psychology* 21 (2): 126–134. doi:10.1111/j.2044-8279.1951.tb02776.x.

Georgiou, S. N. 2008. "Beliefs of Experienced and Novice Teachers about Achievement." *Educational Psychology* 28 (2): 119–131. doi:10.1080/01443410701468716.

Gipps, C. 1999. "Socio-Cultural Aspects of Assessment." *Review of Research in Education* 24: 355–392. http://rre.sagepub.com/content/24/1/355.

Giroux, H. A. 1983. "Theories of Reproduction and Resistance in the New Sociology of Education: A Critical Analysis." *Harvard Educational Review* 53 (3): 257–293. http://hepg.metapress.com/content/A67X4U33G7682734.

Hartog, P., and E. C. Rhodes. 1936. *An Examination of Examinations*. London: Macmillan.

Kain, D. L. 1996. "Looking Beneath the Surface: Teacher Collaboration Through the Lens of Grading Practices." *Teachers College Record* 97 (4): 569–585.

Kalthoff, H. 1997. *Wohlerzogenheit. Eine Ethnographie deutscher Internatsschulen* [Well-Behaved. An Ethnographic Study of German Boarding Schools]. Frankfurt: Campus.

Kalthoff, H. 2006. "Doing/Undoing Class in exklusiven Internatsschulen." In *Soziale Reproduktion edited by W Georg*, 93–122. Konstanz: UVK.

Kingston, P. W. 2001. "The Unfulfilled Promise of Cultural Capital Theory." *Sociology of Education Extra Issue Currents of Thought Sociology of Education at the Dawn of the 21st Century* 74: 88–99. http://www.jstor.org/stable/2673255.

McHoul, A. W. 1985. "Two Aspects of Classroom Interaction, Turn-taking and Correction." *Australian Journal of Human Communication Disorders* 13 (1): 53–64.

McHoul, A. W. 1990. "The Organization of Repair in Classroom Talk." *Language in Society* 19 (3): 349–377. doi:10.1017/S004740450001455X.

McMillan, J. H. 2003. "Understanding and Improving Teachers' Classroom Assessment Decision Making." *Educational Measurement Issues & Practice* 22 (4): 34–43. doi:10.1111/j.1745-3992.2003.tb00142.x.

Moss, P. A., B. J. Girard, and L. C. Haniford. 2006. "Validity in Educational Assessment." *Review of Research in Education* 30: 109–162. doi:10.3102/0091732X030001109.

Pryor, J., and H. Torrance. 2000. "Questioning the Three Bears: The Social Construction of Assessment in the Classroom." In *Assessment. Social Practice and Social Product*, edited by A. Filer, 110–128. London: Routledge: Falmer.

Putwain, D. W. 2011. "How Is Examination Stress Experienced by Secondary Students Preparing for Their General Certificate of Secondary Education Examinations and How Can It Be Explained?" *International Journal of Qualitative Studies in Education* 24 (6): 717–731. doi:10.1080/09518398.2010.529840.

Shay, B. S. 2004. "The Assessment of Complex Performance: A Socially Situated Interpretive Act." *Harvard Educational Review* 74 (3): 307–329. http://search.proquest.com/docview/212263961?accountid=14632.

Shepard, L. A. 2000. "The Role of Assessment in a Learning Culture." *Educational Researcher* 29 (7): 4–14. http://edr.sagepub.com/content/29/7/4.

Starch, D., and E. C. Elliot. 1912. "Reliability of the Grading of High School Work in English." *The School Review* 20 (7): 442–457. doi:10.1086/435971.

Terhart, E. 2011. "Die Beurteilung von Schülern als Aufgabe des Lehrers [The assessment of pupils as the teacher's task]." In: *Handbuch der Forschung zum Lehrerberuf*, edited by E. Terhart, H. Bennewitz and M. Rothland, 699–717. Muenster: Waxmann.

Verkuyten, M. 2000. "School Marks and Teacher′s Accountability to Colleagues." *Discourse Studies* 2 (4): 452–472. doi:10.1177/1461445600002004003.

Weiss, R. 1965. *Zensur und Zeugnis* [Grade and report]. Linz: Haslinger.

Zaborowski, K. U., M. Meier, and G. Breidenstein. 2011. *Leistungsbewertung und Unterricht* [Performance evaluation and education]. Wiesbaden: VS.