How differentiated are assessments in rating performance differences?

A pragmatic method.

Volkhard Fischer[1], Holger Müller[1], Ingo Just[2]

1: Dean‘s office; 2: Dean of Study Affairs, MHH

RIME, Berlin
25. 05. 2013

Medizinische Hochschule Hannover

Introduction

- In 2009 the faculty decided to establish a ranking system for the quality of the modules in the curriculum.
- The system should reward
  - good teaching,
  - the development of more appropriate assessments.
- As main measures were choosen
  - Module evaluation by the students (up to 15 points),
  - Fairness of the assessment (up to 8 points),
  - Size of the module (up to 2 points).

Introduction

- According to the German Licensure Act assessments should differentiate between students with different levels of:
  - Knowledge and
  - Skills.
- Therefore there are two grading systems:
  - In some assessments in the responsibility of the faculty a pass/fail-dichotomy is sufficient.
  - In all other assessments the grades have to be differentiated:

Introduction

Local grade	German name	German description	Translation
1	Sehr gut	Eine hervorragende Leistung	Very good
2	Gut	Erheblich über den durchschnittlichen Anforderungen	Good
3	Befriedigend	In jeder Hinsicht den durchschnittlichen Anforderungen gerecht	Satisfactory
4	Ausreichend	Trotz Mängeln noch den Anforderungen genügend	Sufficient
5	Nicht ausreichend	Wegen erheblicher Mängel nicht den Anforderungen entsprechend	Fail

First considerations concerning the evaluation of the assessment outcome

- Common assumptions about the frequency distribution of grading levels:
  - In the case of a pass/fail-dichotomy
    - a binomial distribution.
  - In the case of a grading system there are several possible solutions:
    - Political-based (e.g. ECTS-Grades)
    - Science-based (e.g. Gaussian distribution)

- Are there any standardized solutions to evaluate a given assessment concerning these assumptions?

A possible solution

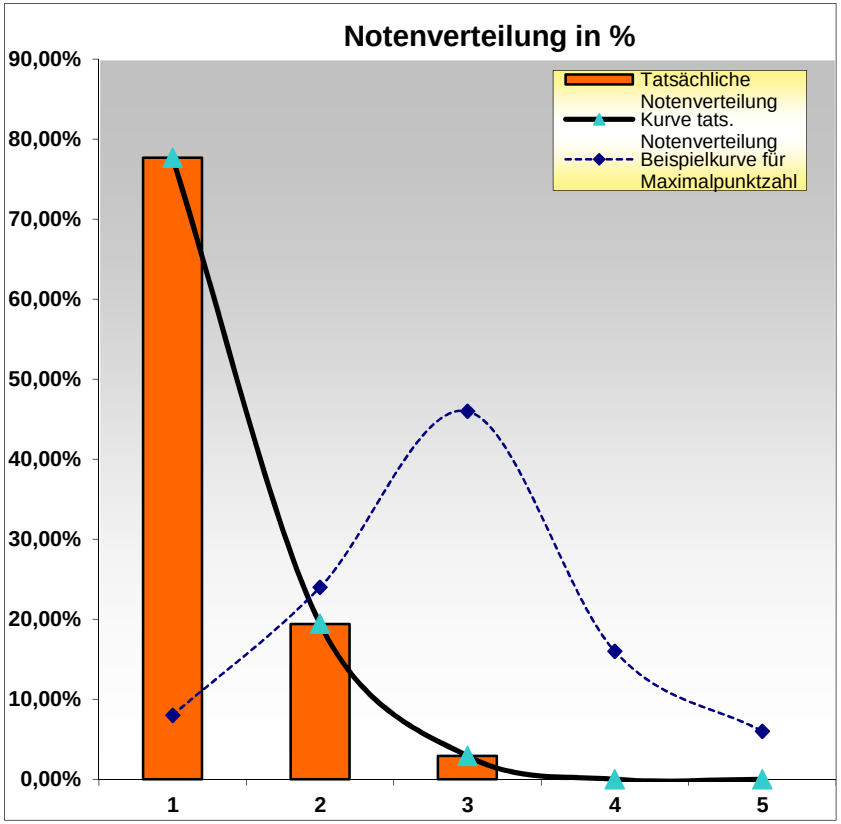
- Kolmogorov-Smirnov-Test for one sample
  - The test is capable to compare an empirical distribution with any given one
    - ECTS
    - Gaussian
  - The test compares the fit of the whole distribution not only deviances of the mean rank (e.g. Wilcoxon-Test)
  - The test can be conducted by standard software or manually
  - There are more effective alternatives (e.g. Anderson-Darling-Test)
  - Nearly all assessments of two academic years deviated significantly from a Gaussian distribution.

The pragmatic solution

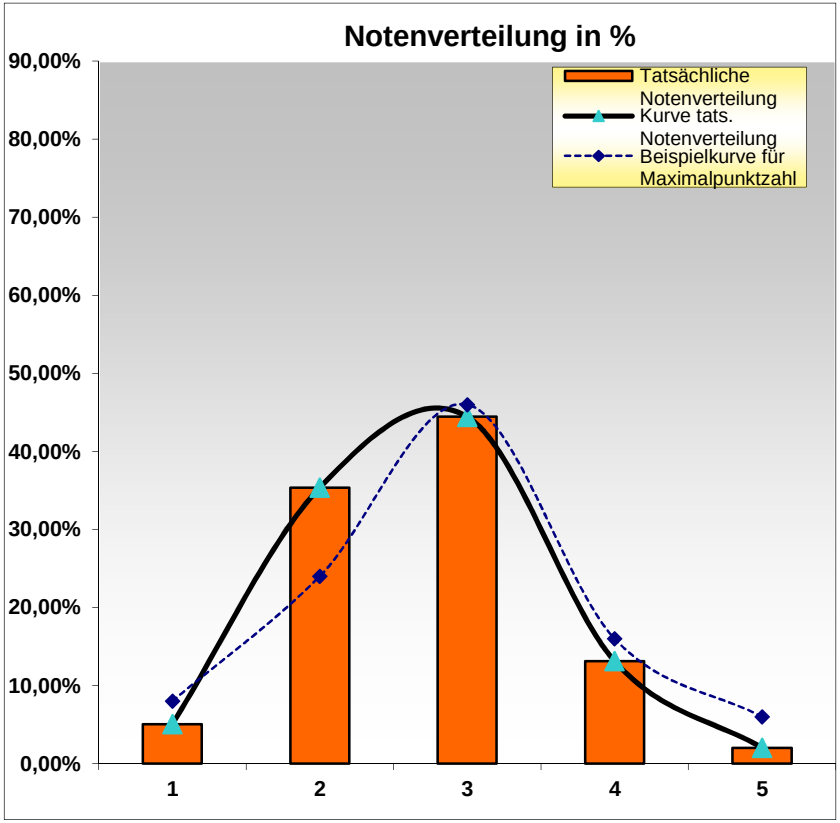
- An algorithm that examines three criteria:
  - The range of the assigned grades
    - Only two grades were assigned => „+0“ points
    - Three grades were assigned => „+2“ points
    - Four or more grades were assigned => „+4“ points
  - The mode of the distribution of the grades
    - There is no pronounced mode => „+0“ points
    - Mode in the boundary area => „+1“ point
    - Mode in the center and pronounced => „+2“ points
  - The proportion of „good“ to „bad“ grades
    - A „good shaped distribution“=> „+1“ point
    - More „good“ than „bad“ grades => „+1“ point
    - More „bad“ than „good“ grades => „-1“ point

Six examples

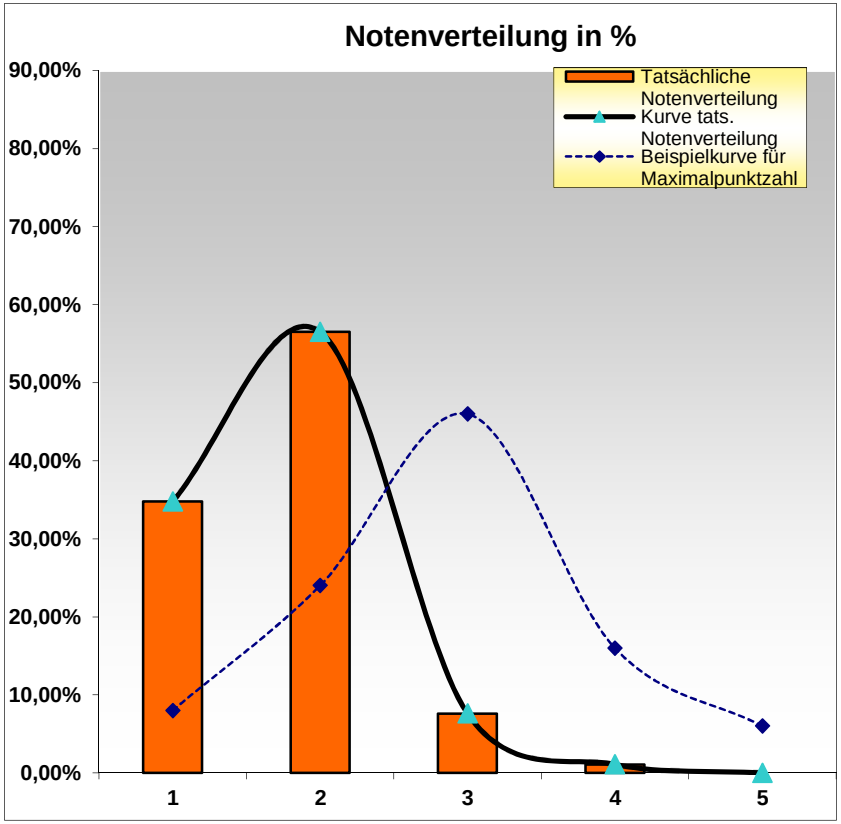
Only two grades were assigned => 0 points



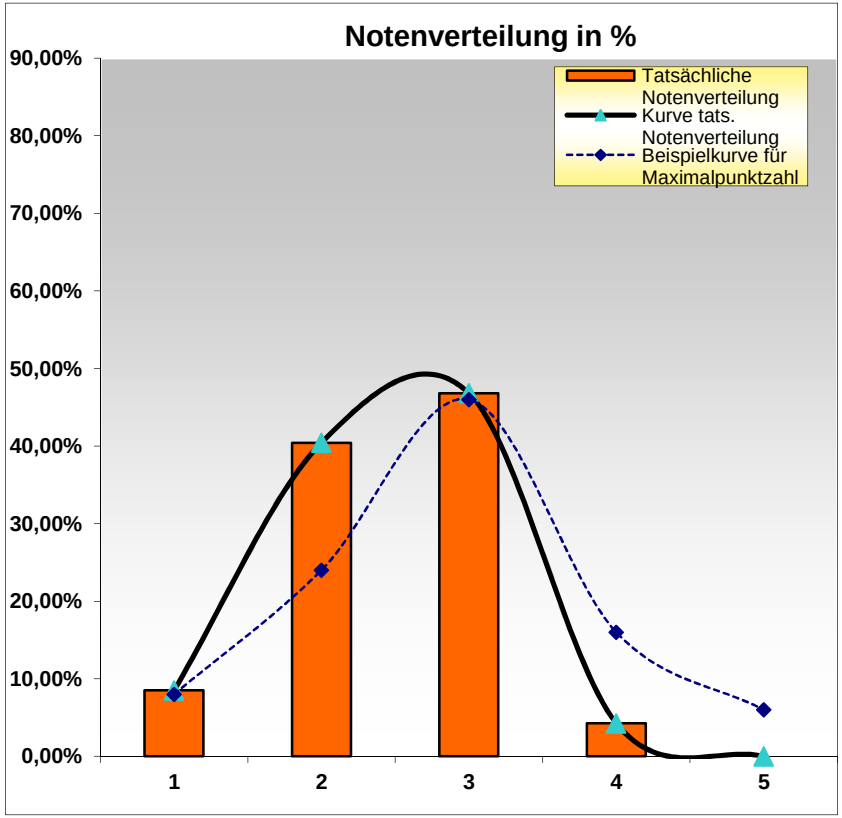
Four assigned grades; salient mode; more „good“ than „bad“ grades => 8 points



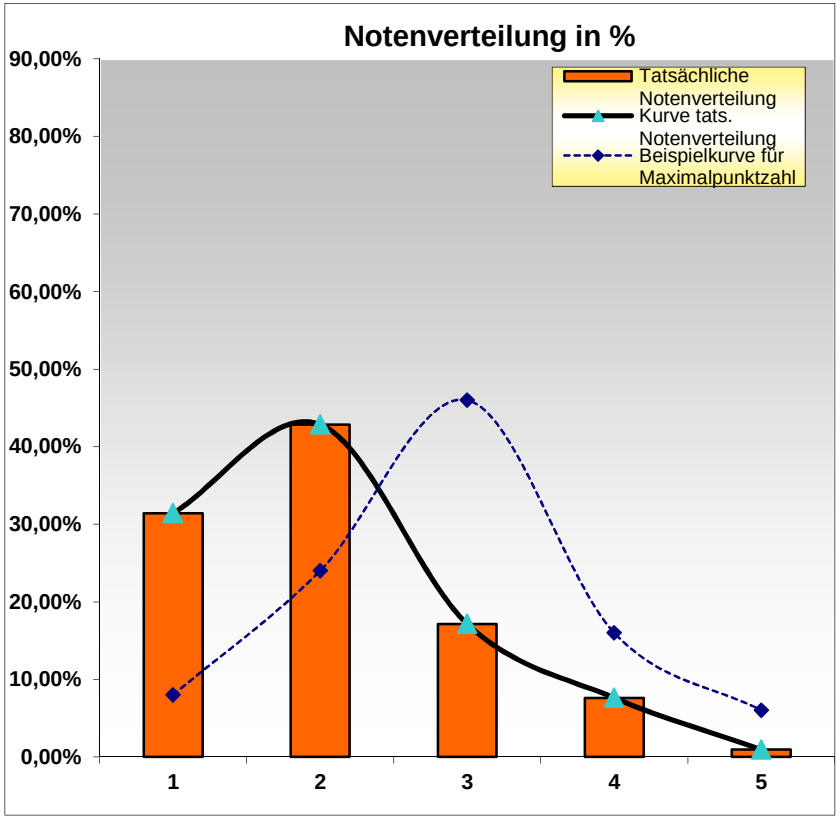
Three assigned grades; narrow distribution with salient mode; more „good“ than „bad“ grades => 4 points



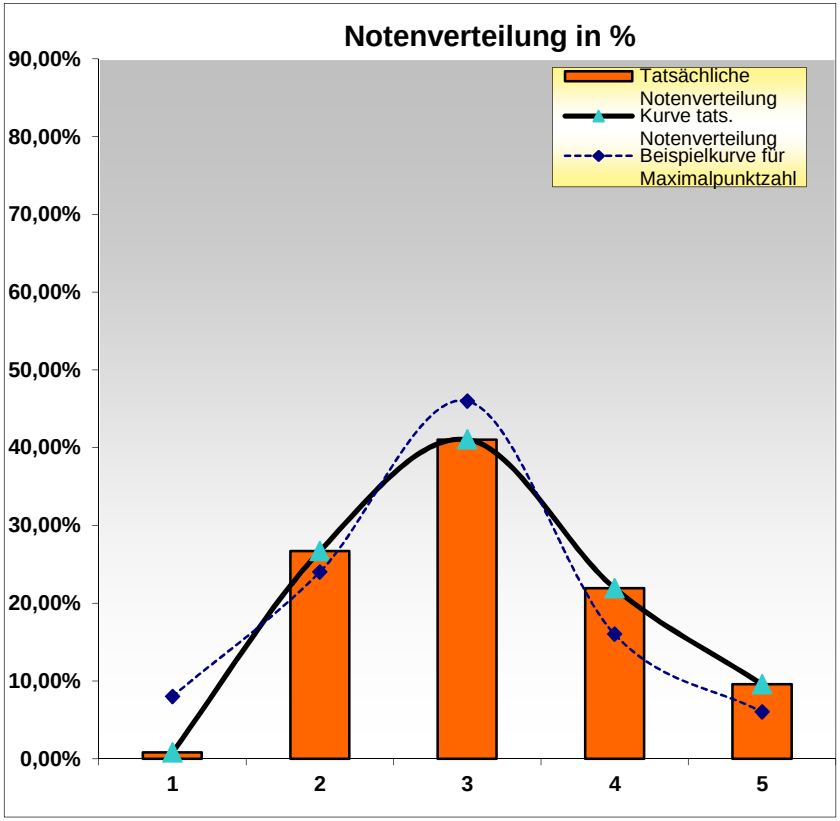
Three assigned grades; narrow distribution with salient mode; mode in the middle of the distribution; more „good“ than „bad“ grades => 5 points



Four assigned grades; wide distribution with salient mode; more „good“ than „bad“ grades => 6 points



Four assigned grades; wide distribution with salient mode in the middle; equal number of„good“ and „bad“ grades => 7 points



Empirical results

- There is only a marginal discrepancy between the academic year 2010/11 and the academic year 2011/12

- Some faculty members were able to optimise their assessments, but a significant proportion clings to their mediocre designs

points	percentages 2010/11	percentages 2011/12
0	15,76%	16,77%
1	0,00%	0,00%
2	15,15%	13,17%
3	0,61%	0,00%
4	24,85%	23,95%
5	7,88%	4,79%
6	29,09%	31,74%
7	2,42%	1,80%
8	4,24%	7,78%

Empirical results

- The algorithm can be used to evaluate all sorts of assessments with at least five grades.
- Many members of the faculty missed that there‘s an algorithm for examining the assessment quality.

Political reactions

- Some stakeholders claim that exams became tougher.
- But there is no evidence for this opinion.
- The discussion about criteria for good assessments has been intensified:
  - An increased interest in formulating adequate questions.
  - Some members of the faculty argue against the use of a grading system.
  - There is a debate about the reward system.
  - But there is (until now) no discussion of
    - appropriate assessment tools,
    - standard setting methods,
    - psychometric theories.

Danke für Ihre Aufmerksamkeit!

Medizinische Hochschule Hannover