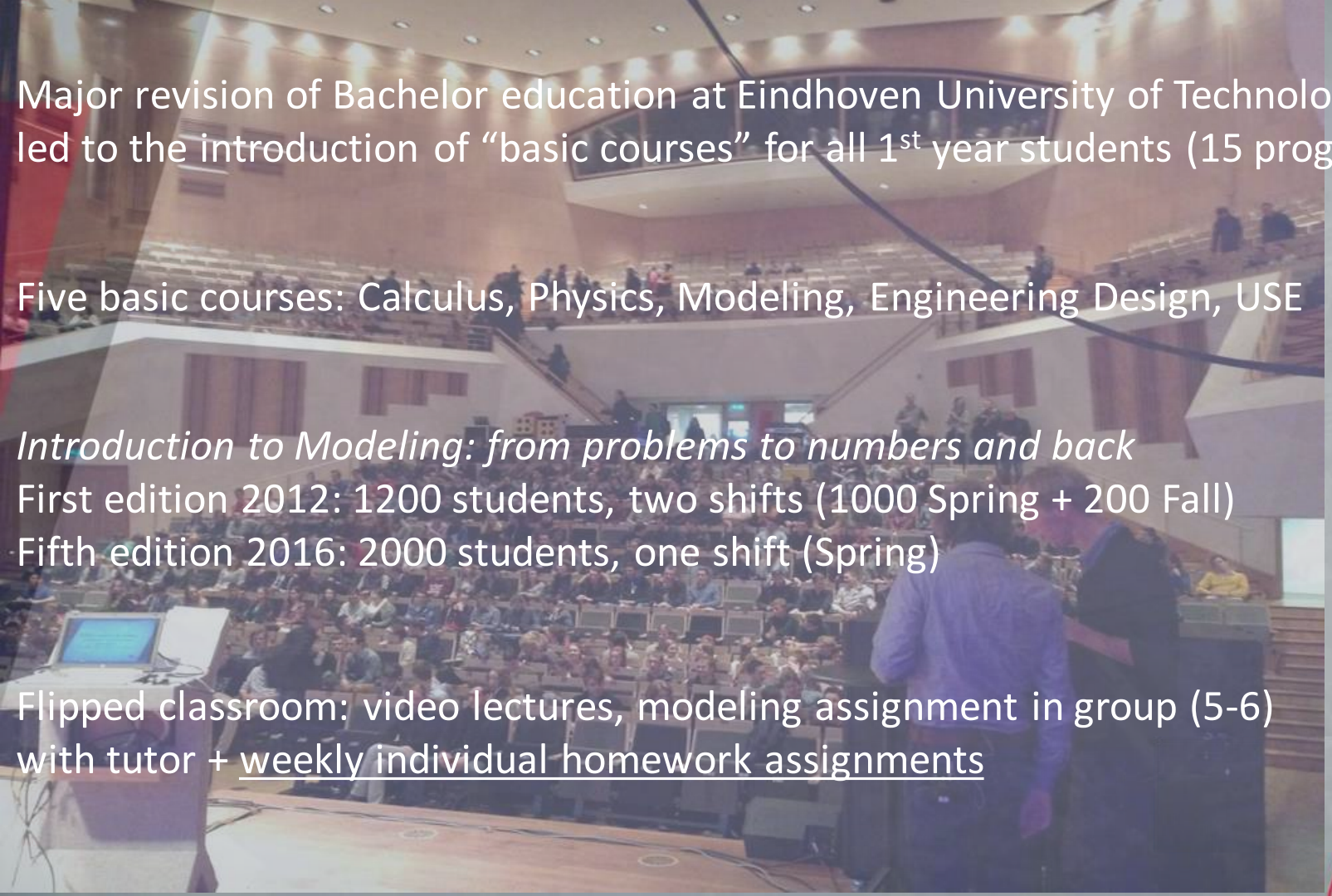
Peer ranking & feedback for the masses

Experiences from the TU/e's Basic Course on Modelling

Tijn Borghuis
Dep. of Industrial Engineering & Innovation Sciences
Eindhoven University of Technology

CDIO Regional Meeting, Delft 26/01 2016

Muziekgebouw
03/02/2015



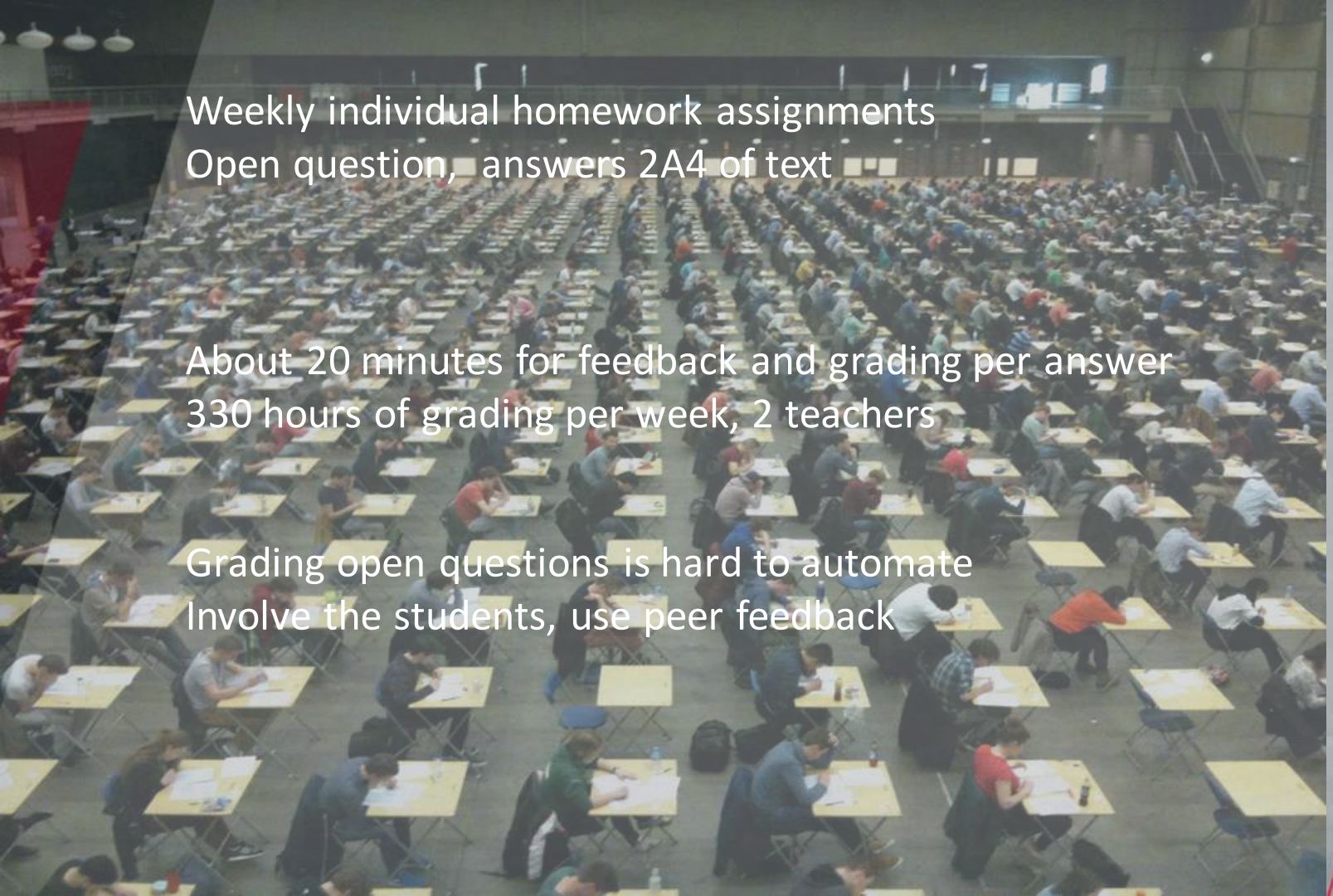
Major revision of Bachelor education at Eindhoven University of Technology led to the introduction of “basic courses” for all 1st year students (15 programs)

Five basic courses: Calculus, Physics, Modeling, Engineering Design, USE

Introduction to Modeling: from problems to numbers and back
First edition 2012: 1200 students, two shifts (1000 Spring + 200 Fall)
Fifth edition 2016: 2000 students, one shift (Spring)

Flipped classroom: video lectures, modeling assignment in group (5-6) with tutor + weekly individual homework assignments

Beursgebouw
11/04/2015



Weekly individual homework assignments
Open question, answers 2A4 of text

About 20 minutes for feedback and grading per answer
330 hours of grading per week, 2 teachers

Grading open questions is hard to automate
Involve the students, use peer feedback

Kees van Overveld
(1957 - 2014)

Tom Verhoeff

At first sight: *peer reviewing* students review each other's work using a protocol that guarantees anonymity

Drawbacks

1. for open-ended assignments, students can be expected to form an opinion ("quite good"), making this opinion quantitative (grade 1-10) is asking too much
2. not all students can be expected to be equally competent reviewers; can be mitigated by having every work reviewed by many students (averaging out); but unrealistic to have students review more than a couple of works a week

By moving to *peer-ranking* comparative ranking in the context of peer review (Allain, Abbott, Deardorff 2006) drawback 1 can be partially addressed (ultimately student works need a grade).

At second sight: *peer ranking* students review each other's work using a protocol that guarantees anonymity

Drawbacks

2. not all students can be expected to be equally competent reviewers; can be mitigated by having every work reviewed by many students (averaging out); but unrealistic to have students review more than a couple of works a week

How can peer-ranking be used, taking differences in students' reviewing competences into account, in order to obtain absolute marks in assessments?

By moving to *self-consistent peer-ranking,* loosely based on HITS algorithm (Kleinberg 1999) used for self-consistent ranking of scientific citations

Proposed approach: self-consistent peer ranking

Compare to the problem Google solves when by page ranking

- a webpage is *good* if many pages link to it;
- not every link should contribute equally to the goodness of a web page;
- a link from a *good* webpage should contribute more;
- this gives a cyclic definition of what constitutes `*good*' for web pages

Proposed approach: self-consistent peer ranking

In the case of peer ranking the reasoning goes as follows:

- a student's work is *good* if peers have a high esteem of it;
- not every peer's opinion should contribute equally to the goodness of a work;
- the contribution of a *competent* peer should contribute more;
- this gives a cyclic definition of what constitutes 'good' (for works) and 'competent' for reviewers

With an iterative algorithm (Van Overveld, Verhoeff, 2013) the differences between students' ranking competence are estimated, and used to compute a weighted final rank score of the works.

Teachers review the highest and lowest ranking work (and perhaps more for increased reliability) in a cluster to establish the absolute marks; marks of works not reviewed by teachers are found by interpolation

Implementation using peach (http://peach3.nl):

Open-source web-based system to handle work submitted for assignments

Supports manual and automated feedback

Stable, in routine use for many courses at TU/e and beyond, since 2001

Fall 2012: Supports double-blind peer reviewing

- Ranking
- Grading
- Reporting

Protocol: self-consistent peer ranking

Homework assignments (weekly)

- open question;
- published rubrics for assessment (binary);
- students rank the works of 5 others using rubrics;
- grades for homework assignments count towards course grade

Direct test of modeling skills

All distribution and processing of documents and student rankings handled automatically by Peach

Assumptions: self-consistent peer ranking

- students are able to recognize quality, also if it is above their own level;
- students raise their level to that of superior peers;
- students have an intrinsic “performance level” and “ranking level”;
- wisdom of the crowds will lead to statistical convergence

Model studies (simulations) showed that self-consistent peer ranking was possible

Practice: self-consistent peer ranking

Spring 2013 (900+ students in 26 clusters of 35-40 students)

Assumptions didn’t hold

- correlation of student scores with teacher samples was disappointing;
- statistical convergence over the weeks was slower than expected

Students have low trust in the results

- difficult to accept the possibility that “inferior” students judge their work;
- the fact that the “noise” on the grade is no greater than when different teachers would grade their work is not recognized as an argument

Alternative: peer feedback

Homework assignments (weekly)

- open question elaboration;
- published example elaboration;
- students give free text feedback on the work of 3 others;
- recipients of feedback give a “like” if they felt they learned something from it;
- grade for feedback (O(# likes received))

Test of feedback skills, modeling skills assessed in separate traditional exam

All distribution and processing of documents and student likes handled automatically by Peach

Assumptions: peer feedback

- students are willing to do homework, if only to get access to feedback process;
- Students are able to judge when feedback is instructive;
- students have an intrinsic “feedback quality”;
- wisdom of the crowds will lead to statistical convergence

Practice: peer feedback

Spring 2014 (1000+ students)

Assumptions mostly hold but

- not all students willing to do homework to take part in feedback (uploading random pics, promises to like,…) ➔ introduction of a “flagging” mechanism

Students trust the principle

- they accept that a student is able to judge what is instructive for him/her;
- they respect each other’s judgements

But want guarantees for practice

- they don’t accept likes can be “lost” when a recipient of their feedback forgets to log in to give likes (separate round in the protocol)
- they elaborate proposals (dislikes, checking user activity logs,…)

Peer ranking and feedback for the masses

Technology
- Peach just works
- logistics and scoring handled smoothly for 1000+ students

Concepts
- Self-consistent peer feedback needs work (when do assumptions hold?)
- Peer feedback is sound

Culture, the main issue
- maturity of the students (1st year students didn't accept peers grading their work)
- formative vs summative (points motivate, but can also elicit unsocial behavior)
- anonymity (sometimes leads to "madness" rather than "wisdom" of the crowds)

References

Allain, R., Abbott, D., and Deardorff, D. (2006). Using per ranking to enhance student writing. *Physics Education*, 41(3): 255-258

Kleinberg, J. (1999). Authoritative soureces in a hyperlinked environment. Journal of the ACM, 46(5): 604-632.

Van Overveld, K., Verhoeff, T. (2013). Self-consistent Peer Ranking for Assessing Student Work – Dealing with Large Populations. In *Proceedings of the 5th International Conference on Computer Supported Education*, 399-404.