Project Report **EvaluaTUM**

Team   Judith Paripovic
Pascal Resch
Paul Sieber
Stefan Engels
Stefan Lehner

Tutor   Konstantin Riedl

Supervisor   Prof. Dr. Annette Noschka-Roos

EvaluaTUM

Preface by the Supervisor

Prof. Dr. Annette Noschka-Roos

There is a magic in every beginning. This well-known quotation from Hermann Hesse always comes to mind in supervising different project teams of the TUM: Junge Akademie (TUMJA). It's characteristic of magical events that they could not be explained in full, but there are some, even essential, ingredients helping to make the magic. First of all the concept of TUMJA has to be mentioned. Choosing and combining students from different domains or faculties brings together different perspectives, different methods for solving problems in an interdisciplinary manner. The students themselves decide on the problems which they want to solve during their two-year projects concerning social issues: projects with subjects like "science goes public" or with health issues within the university or even within the community. This procedure can also be considered an important source of the magic: choosing a problem and starting a project establishes commitment among the students.

In this case the team of the project EvaluaTUM wanted to improve the current evaluation system with a central question in mind: could it be useful to develop the evaluation instrument in such a way that it improves the communication between teachers and learners by direct feedback? And does this help to improve the seminars, lectures and courses, too? Both a simple and a complicated question including tricky and complex issues: different faculties, different domains, different methods for teaching, different personalities on teaching and learning sides, different experiences and many other factors in this multifaceted field.

Choosing the top-down approach for their evaluation project the team of EvaluaTUM convinced two faculties to take part in this project. They developed evaluation instruments in each case with specific questions concerning alternative response times during the lecture, the students participating in the evaluation, the length of the questionnaire, and the basic tools (paper/pencil vs. digital). They were theoretically convinced that these factors, among others, have an impact on the evaluation concerning the feedback quality. Would it have been better to look for examples of best practice using the common instruments for systematically identifying the leading factors for success, a bottom-up analysis? But this would be a time-consuming and methodologically sophisticated approach, so the students used this more suitable one. They worked in an impressive way independently, enthusiastically and with discipline obliged by the fact that some of them were located abroad as guest students. Confronted with a difficult task, the students did a great job. Even though the results cannot be generalized, they are helpful for thinking about previous practice in relation to the current evaluation system. They proved the impact of some factors they had presumed: the length of the evaluation questionnaire, the possibility of answering open questions, the effect of a provided timeframe.

In the end of the evaluation project, another essential ingredient of the magic start is worth mentioning: the dream and the hope of solving an identified problem. It isn't solved but the results show that it was highly worthwhile to consider it seriously. ∎

No Excellence without Evaluation

It might seem evident that lectures in general need constant improvement. This also applies to the best lectures. The Technical University of Munich is often said to be excellent – not only in research – but also in teaching. This excellent teaching needs evaluation – constant, outstanding evaluation. Considering all the excellence of TUM, the question might arise why their professors – the smartest and most educated people – need evaluation for teaching?

Obviously, for the last 150 years, professors have been teaching at TUM. Also obvious, not only professors teach at universities. Many lecturers are doctoral students and thus are right at the beginning of their careers. This implies that they still need and want to improve their teaching skills. Although professors are already experienced lecturers and highly educated, there are reasons why a steady evaluation is necessary.

A professor's job consists mainly of two aspects: research and teaching. Nevertheless most of them see themselves mainly as scientists – not as lecturers – and thus focus more on the former of the two tasks. Therefore training of lecturers is essential.

Unfortunately some lecturers ignore the evaluation process. Luckily they are a minority. Their point of view is: "I am the expert, so I know best what and how to teach." Within this sentence, they ignore several points. Firstly, many students do not comprehend the content as fast as professors. Hence, the students' input is essential. As nobody is perfect, improvement is always possible, especially for teaching students. Perhaps the most important point has nothing to do with the lecturers themselves. Nowadays, everything develops fast. New teaching strategies develop rapidly but not every strategy suits every content of teaching. As students listen to many lectures, they see a lot of different teaching approaches every day. Therefore, they can give valuable feedback. Moreover, not only the methods of teaching change, but also the contents. From this it is evident that even the best lecture concept has to be improved and adapted continuously. Especially, at a so-called excellent university. Therefore it always needs outstanding feedback and evaluation.

Contrariwise, the feedback rates at TUM show that there is still room for improvement in the field of lecture evaluation. The members of team EvaluaTUM tried to tackle this potential in their research. ■

Low Feedback Rates are not the Students' Fault!

Every semester, lectures at TUM are evaluated. The aim is clear – improvement of the classes. Unfortunately, many lecturers and administrative staff complain about incredibly low feedback rates. They highlight the students' low motivation and disinterest. But NO. The low feedback rates are NOT the students' fault!

First of all, we – a group of students – need to clarify that none of our colleagues ever refused to give feedback. Whenever lecturers ask how to improve their teaching, they will get the student's honest opinion about their course. Criticizing bad courses makes a student feel less helpless, hoping that the second half of the semester will get better. Similarly, replying positively to a class means possibly suggesting the lecturer for a teaching award. Motivation from the students' perspective does exist and is even quite high! Therefore, the low participation must have different reasons. It looks as if the real problem is a structural one!

Also technical issues need to be overcome: Instead of clear and structured processes, online evaluations are sent via email, and students also receive many other emails every day. Sometimes, one mail per evaluation form is required, leading to a huge number of emails in the inbox. Of course, the motivation to click all ten or twelve links is low. One link for every lecture, one link for every exercise, and one link for the labs. These anonymously sent requests for feedback make a student think: "Spend the next two hours with 'EvaSys'?!" But emails containing the evaluation requests are patient – they can wait. If necessary, forever… NOT on purpose. It just happens because of the inbox's nature.

If students remember to fill out the survey, they get a bunch of questions with scales from one to five. After half a page, every student starts to tick intuitively. As the questionnaire seems never ending, the second page is ticked with less thought. Attentive students notice that question 29.6 on page 4 is the same as question 2.5. Also the layout – grey and the newest design from the 90s does not help to keep the initial motivation. After the pages of single-choice questions, the students' motivation is finally gone. The problem is that the last page contains the most helpful part for the lecturer: The open-response questions.

The motivation of the lecturers sometimes seems to be similarly high as the students' after five pages. As most of the lecturers do not give time for filling out the online questionnaire during the class, it feels like giving feedback is a waste of time. Some do not know the exact evaluation deadline! Additionally lecturers often do not discuss the results. Maybe the results are handed out too late – but motivated lecturers could even discuss the previous year's results in order to show their interest in improvement.

To conclude, accusing the students for low feedback rates might be a little bit too easy. But in reality, all lecturers can motivate their classes by allocating time for the evaluation itself and reviewing the results together with the students. In addition, the administration could try nicer layouts as well as shorter questionnaires with more focus on free-response questions. To show the influence of these aspects, EvaluaTUM conducted some research. ∎

The Influence of Three Different Factors on Lecture Evaluation at TUM

Abstract

To ensure high teaching quality, a good and constantly adapted lecture evaluation is necessary. Technical University of Munich (TUM) has therefore developed a complex evaluation system within the scope of their quality management. Considering the existing system, this report shows that small changes in variables can improve the return rate and feedback quality of lecture evaluations. Analysed independent variables are the time slot given to complete the questionnaires during lectures, the chosen evaluation media (online vs. paper based) and a modification of the questionnaire in terms of length and question type. Therefore, we implement three case studies to validate the following three hypothesis:

1. Giving students time to fill out the questionnaire in the middle of the lecture leads to significantly higher feedback rates compared to not motivating the students during the lecture.
2. Online evaluations increase the return rate in comparison to paper-based evaluation if both are conducted similarly.
3. A shortened evaluation questionnaire with focus on open questions contains more feedback than a questionnaire with focus on single choice questions.

The collected data shows in the first place that the middle of the lecture seems to be the best point of time for maximizing the feedback rate. Secondly, online-based evaluation does not significantly influence the response rates, although more students are reached, i.e. students who do not visit the lecture that specific day can also participate. Thirdly, a shorter questionnaire with focus on open questions improves the feedback quality.

From the latter two points one can derive that small changes in the current lecture evaluation system at the TUM may lead to significantly improved lecture evaluations. Many instances of TUM have already reported their interest to our results in order to improve the evaluation system. One faculty already initiated the rethinking process of their evaluation sheets due to this project.

Introduction

Motivation

Universities are places where research meets higher education and thus lecturers must manage both the scientific research of highly

complex contents as well as giving lessons to students. In teaching, they have to consider various aspects: e.g. keeping lectures up-to-date, breaking down complex topics, and being available for their students. To achieve these tasks and to be able to offer an excellent teaching at the university, feedback loops are essential. Feedback from the students to the lecturer is mostly given by evaluation questionnaires. However, with a high variety of subject areas and lecture types, the range of evaluation means and forms is very wide. There are e.g. classical lectures, practical courses or seminars given by a single or multiple lecturers. When creating and implementing a questionnaire several questions occur:

- Which type of question is optimal for the goal of this questionnaire?
- What is the best ratio between open questions and single-choice questions?
- What is the optimal length of a questionnaire?
- How can one achieve a reasonable return rate?
- How can one ensure a comparability of the lecture feedback?

Besides these uncertainties in the questionnaire development, especially at larger universities, the implementation of general feedback loops is difficult and can be interwoven with the whole system. Due to this high complexity, there is always room for (small) improvements of these feedback-loops. A survey among student representatives at the Technical University of Munich (TUM) confirms improvement potential (Fig. 1). The following report presents the influence of selected levers to increase the feedback quality.

Current System at TUM
At the paper- and online-based questionnaires are typically used including demographic and general questions as well as open questions with room for individual feedback. The evaluation process at TUM consists of three steps. The base is built by the evaluation of courses. Additionally, the degree programs and the entire faculties are evaluated.

To perform the evaluation process, the faculties have responsible employees for quality management. They are supported by the "TUM Center for Study and Teaching - Quality Management, Legal Division and Communications" at university level. A third player within this process is "ProLehre". They help lecturers to interpret

the results and develop plans to improve teaching. Additionally, they perform Teaching Analysis Polls in classes, where the number of students is so low that evaluation forms are not anonymous anymore. [30]
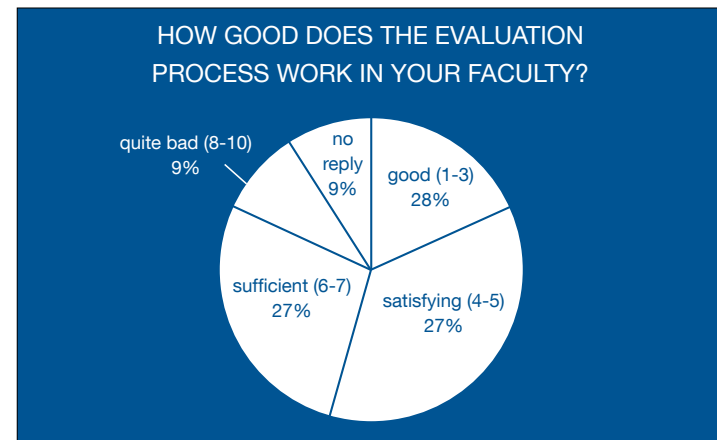

Fig. 1: Response of the students' representatives of 11 out of 21 different faculties to evaluate the lecture evaluation quality (survey of EvaluaTUM, [7])

The development of the evaluation forms is each faculty's responsibility. However, there are approximately eight mandatory questions, which each questionnaire must contain. They are important for comparability. The class evaluation takes place either online or paper based with evaluation forms from EvaSys. Some faculties even use a hybrid system. A survey among the student representatives of different faculties shows the high variation of evaluation media used (Fig. 2). Similarly, course evaluations take place at different points of time within the semester, but also within a lecture. The above mentioned survey points out that some faculties evaluate during the lectures, whereas others rely on the students' evaluation at home (Fig. 3).

In addition to the system at TUM, we were able to experience personally the evaluation systems at Kungliga Tekniska Högskola (KTH) in Stockholm as well as the Eidgenössische Technische Hochschule in Zurich. Derived from all these possible realizations of evaluation and the current implementation at TUM, our main

goal of this research is to optimize the evaluation questionnaires at the TUM for better feedback and therefore a further improvement of the teaching.
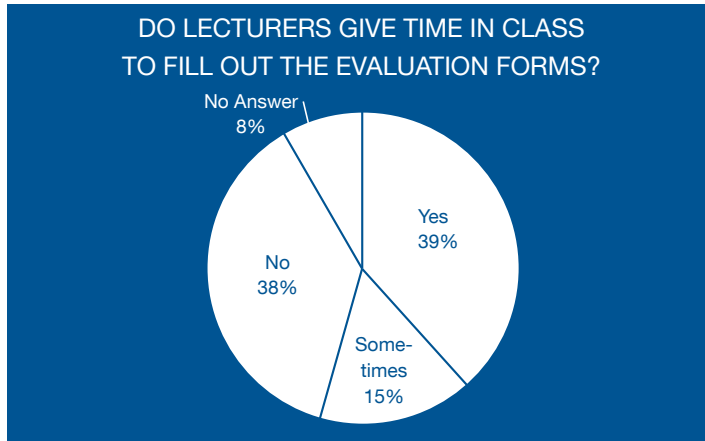

Fig. 2: Response of the students' representatives of 11 out of 21 different faculties to evaluate if the lecturers give time for the evaluation during the lecture (survey of EvaluaTUM, [7])
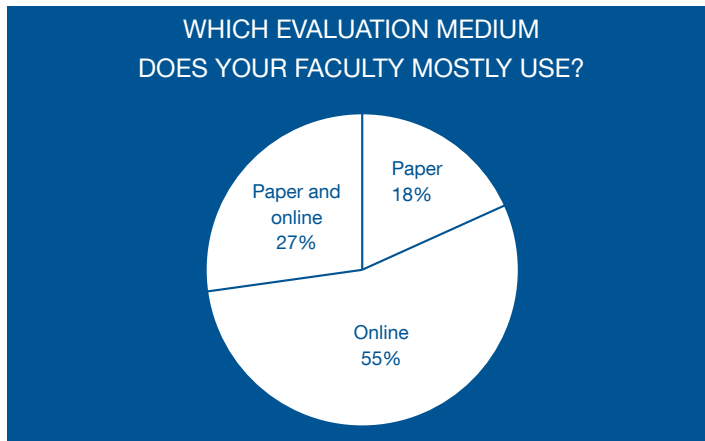

Fig. 3: Response of the students' representatives of 11 out of 21 different faculties to evaluate which medium the faculty uses for the lecture evaluation (survey of EvaluaTUM, [7])

Problem Formulation

To obtain significant evaluation results, a large number of students should participate in an evaluation (Rantanen 2013). In order to increase the occasionally low feedback rates, modifications in the current evaluation method are necessary. Our overall goal can be described as the suggestion of changes, resulting in higher response rates and a more quality feedback outcome in the lecturers' view in order to maintain the full feedback loop. Considering the complexity of the whole lecture evaluation process stated in the first part of the introduction, the leeway for changes is very limited. This leads to the project's research scope: Investigating possible modifications of lecture evaluation within the current system at TUM and specifying their impact.

Literature Review

The following section gives an overview about important results that have been presented in the literature.

Reliability and Information Value of Evaluations

Standardized evaluation questionnaires require a high reliability of the measurement. Even if the participants of an event give different evaluations, the results are stable on average over the survey methods, measurement dates and groups of people [20,24]. As a long-term study (6024 courses, 195 lecturers, 31 courses of study, 13 years) has shown, the results are very stable even if different students evaluate the courses several times over a longer period of time [17,18]. This shows that the evaluations of 74% of the lecturers changed only slightly. Evaluation results are stable even after completion of evaluations. An analysis of 6 individual studies shows that the variances in evaluations by current and former students are identical at about 48% [8]. A further study, in which 100 events were evaluated in a contemporary way and years later, shows a variance of 69% [19].

Implementation and Response Rate

In order to optimize the significance of evaluation results, as many students as possible should participate in an evaluation. About 15 completed evaluation questionnaires are needed to make an evaluation result generalizable [24]. Rinderman and Schofield [25] as well as Kreiter and Lakshman [11] also stated that the evaluation results of about 15 persons are needed to make generalizable statements about the quality of a course. In general, the participation rate in evaluations is far below 100%: according to studies

between 30% and 70% of the students that take the class [1,3,6]. This also holds for the TUM, where according to our first study [7], the response rates vary from 3% to 90% among different disciplines and module levels.

Students are more likely to participate in evaluations if they believe that the lecturers will implement the evaluation results in their lectures [4]. As one study shows, an event should be evaluated by only 15 randomly selected participants to make the results more meaningful [11].

The used media (online vs. paper questionnaires) has almost no influence on the evaluation results [6,12,28,30]. At TUM, faculties evaluating with paper seem to have higher response rates than the ones using online evaluation [7].The time of evaluation during the semester also plays only a minor role. Studies show minimal positive effects when an evaluation was carried out in the middle of the semester [5,14].

Interpretation of Evaluation Results
In principle, evaluation results are only guidelines for the lecturers. They also lead to a whole range of emotions of the lecturers (e.g. discouraged, frustrated, angry, motivated, satisfied). Often it is the case that free comments contain hurtful statements [9,15]. Evaluation results do not necessarily lead to an overall improvement of an event. As a study from Germany (3122 evaluation sheets, 12 lecturers, 4 semesters) shows, the quality of a course does not improve despite repeated evaluation [10]. With additional didactic counselling and further education offers, lecturers receive the necessary background knowledge to be able to implement evaluation results accordingly.

Open Feedback Questions
A less frequently used method to design questionnaires is the use of open response questions. For these types of evaluation forms, only little research has been done [2]. Nevertheless, open questions might be an important aspect to design evaluation forms as well as to specify improvement possibilities. This paragraph summarizes the most relevant aspects. The percentage of students responding to open questions varies significantly from 10-70% [2,22]. The length of the responses depends also on the medium: Online questionnaires lead to greater participation in open questions and to longer comments [27]. A higher percentage of comments is generated by short questionnaires [27]. Oliver et al. [22]

as well as Zimmaro [31] observe that positive comments prevail in comparison to negative comments. Moreover, positive answers tend to be more general, while negative aspects are rather specific [31]. Although lecturers prefer to interpret individual comments, they are more difficult to interpret [13]. According to Alhija and Fresko [2], the students' answers often comment on the courses' conceptual contents and state specific weaknesses of the course context. This might lead to the assumption that open questions are used for more specific, rather critical feedback which cannot be covered by standardized single-choice questionnaires.

Research Question
As the lecture evaluation system at TUM is interconnected with the accreditation process of the single degrees, the overall structure of the evaluation process is rather strict. Thus, the range of possible modifications are limited and only minor changes can be implemented. To obtain more significant results from evaluation at TUM and to increase the response rate within the given frame at TUM, we will focus on three questions:
1. Does the point of time to fill out the questionnaire in the middle of the lecture lead to significantly higher feedback rates compared to not motivating the students during the lecture?
2. Will online evaluations increase the return rate in comparison to paper-based evaluation if both are conducted similarly?
3. Does a shortened evaluation questionnaire with focus on open questions contain more feedback than a questionnaire with focus on single choice questions?

All three research questions imply rather small changes of the current evaluation system, but will nevertheless provide interesting insights in the complexity of evaluation. Also, these three aspects can be decided by the faculties or the lecturers themselves and therefore do not need big changes.

To answer the questions, the project group carries out three case studies. In the following, the methods and results for the three aspects are presented.

Methods
In this section, the scientific approaches for our research are described. Firstly, we briefly explain how we identify potential in the existing evaluation system. This section is followed by three sections describing the methods used for each research question, respectively.

Evaluation of Needs

In a first step a survey is conducted to find needs within the evaluation system. The questionnaire is sent to the student representatives of all faculties and 11 of 21 faculties participated. It includes questions such as the feedback rate and the medium used for evaluation. Additionally, open feedback is collected. To collect further information, interviews with experts such as quality management officers and lecturers are conducted.

Case Studies

1. Effect of Provided Time Slot

In a first study, the aim is to find if there is a correlation between the feedback rate and whether lecturers give time for evaluation in their lecture. Therefore, 17 lectures in the winter term 2019/2020 at the Munich School of Engineering were investigated. The evaluation takes place within two weeks in december. All students get an email with links to the evaluation of all lectures for which they register.

We split the lectures randomly into three equally large groups. Lecturers of classes in the control group are instructed not to give time during their lectures for evaluation. However, they are allowed to motivate their students to evaluate the lecture from home. In the other groups, the lecturers give approximately 10 minutes within one 90 minute lecture. In the first of these two groups the students get time for evaluation in the middle of the lecture, whilst the other group has time for evaluation at the end of the lecture.

For significance testing, we use the Tukey HSD (honestly significance difference) test. This test compares the mean differences of all pairs. Together with the mean differences, we state the p-values. This test assumes that the means of all pairs are the same. The p-value expresses the probability to see our or more extreme data given that the former assumption is true. A small p-value suggests that this assumption does not hold and the means differ. By convention, a p-value less than 0.05 is said to show a statistically significant difference.

2. Impact of the Evaluation Medium

To compare the impact of the evaluation medium used (online vs. paper based evaluation), we use Welch's t-test. This test compares the means of two different groups of different sizes. It assumes that these means are the same. The p-value expresses the probability to see our or more extreme data given that the former assumption is true. A small p-value suggests that this assumption does not hold and the means differ. By convention, a p-value less than 0.05 is said to show a statistically significant difference.

3. Questionnaire Design

For the third study, we compare an existing questionnaire (shown in figure 04) at the department of chemistry at TUM with a new questionnaire designed by our team (depicted in figure 05) in one lecture with 80 students. The current questionnaire consists of a double sided A4-paper with 36 single-choice questions and about one fourth at the end of the second page were filled with two open text questions. The modified questionnaire is also one double sided questionnaire but except for the first question about the study program, the first page contains only the same two open questions. The second page is composed of 12 single choice questions which are considered as essential for a general evaluation according to TUM guidelines or the faculty of chemistry. This leads to a shortened questionnaire with a distinct focus of open-response questions.

In order to obtain unbiased results we hand out the questionnaire versions alternating the seat rows in the lecture hall. We do not make any announcements regarding the different evaluation sheets beforehand to avoid students recognizing the two versions. Almost no students recognized that there have been two kinds of questionnaires.

One week later, we perform a TAP evaluation to get a ground truth for the feedback which the students wanted to communicate. The

Feedback rates depend on the reference participation number. Therefore, we measure it in two ways. Firstly, we compare the number of filled evaluations against the number of students registered for the lecture. As there are students only formally enrolled for a lecture, the ratio of filled forms with registrations for the exams is used as a second measure. Bias arises as some of the students take classes, but write the exam one semester later. Especially modules which run over two or more semesters are a problem. Classes, where this is the case, were not used for the second measure.

TAP is a qualitative method for interim evaluations, during the semester. The TAP is moderated by an external person and the lecturer should leave the lecture hall in order to maintain anonymity. The process can be summarized as follows:

1. **Formation of Groups:** Students are divided into small groups (between 3 and 7 people). Morgan showed, that large group sizes do not contribute to the results [21]
2. **Questioning and Group Discussion:** The moderator poses three questions to the students. The students have 10 minutes to discuss their opinions.
   a. Which aspects of the course do you consider as beneficial for studying?
   b. Which aspects of the course do you consider as an obstacle for studying?
   c. Which suggestions to improve the lecture do you have?
3. **Documentation:** Using key points, the students note their answers. The moderator collects the answers, clarifies ambiguous and misleading statements and reformulates them accordingly. His task is called "Formulating interpretation" [16]
4. **Analysis of the Data:** The moderator determines coding units during the collection of the data. By means of these coding units, the answers are grouped and sorted, after the TAP data collection is completed.
5. **Feedback to the Lecturer:** The first feedback is given by email, promptly after the TAP
6. **Feedback Meeting:** This meeting is conducted according to the "Consultative approach" [23]. The agenda includes the discussion of the lecturers self-reflection and the results of the TAP. The moderator and the lecturer develop together strategies for improvement.

This method has certain advantages: Context-related quality characteristics are captured. Furthermore, the evaluation (TAP) and consultancy is connected. This is very important for effectiveness. Unfortunately, the evaluation results depend on the TAP moderator. Different moderators may weigh opinions subjectively, which results in a non standardized analysis of the answers.

For our research, the procedure is slightly adapted: The students paired in groups and posted their negative and positive aspects on tweedback. The documentation, moderation and data analysis are replaced by "Likes" on tweedback, indicating the relevance of each aspect. This leads to a list of negative and positive aspects as well as improvement suggestions, ranked by importance.

Fig. 4: Current questionnaire for lecture evaluation at the faculty of chemistry at TUM

Fig. 5: Modified questionnaire for lecture evaluation with fewer questions and focus on open questions

Results

In this section, we present the results for all three case studies. We begin with the effect between paper- and online-based evaluation in general to see if the medium used has a significant impact on the number of participants. Secondly, we compare the different groups within online evaluation to assess how the response rate can be

maximized. As a third part, the quality of answers received from two different questionnaires are compared.

First, for the former two studies, we removed two outliers from our data set. Due to very few exam registration, the corresponding response rates exceeded 100% and therefore do not provide a meaningful measure. Thus, a total of 15 lectures have been analyzed. In 4 lectures no time was given to fill out the questionnaire. 5 Lectures provided time around the middle of the lecture and 6 lectures at the end of the lecture. We also collected corresponding numbers of paper evaluation the preceding year.

1. Effect of Provided Time Slot

The provided time for filling out the evaluation questionnaire varies strongly at TUM. We analyze how the different ways of performing online evaluation affect the response rate. The boxplots in Fig. 6 and Fig. 7 present the different response rates measured with respect to lecture registrations and exam registrations respectively. The color and x-axis show the group (no time provided, time provided around the middle of the lecture, time provided at the end of the lecture), whereas the y-axis depicts the corresponding response rate. As described in Methods, we use Tukey's HSD test for significance testing.
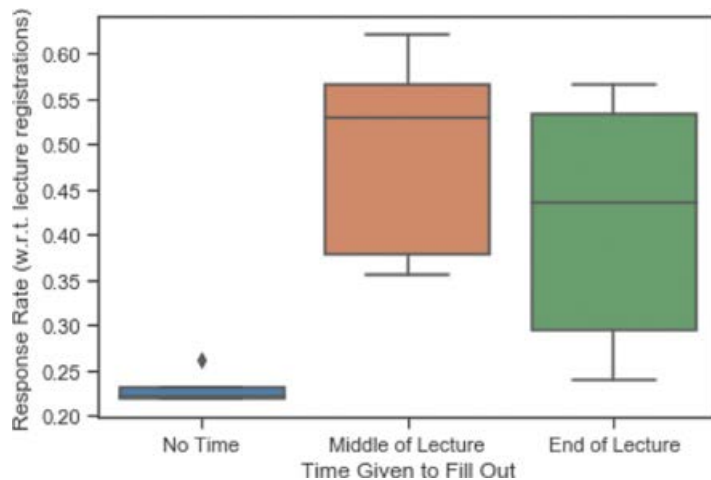
Fig. 6: Response rate with respect to lecture registrations of the investigated lectures, grouped by the time provided to answer the questionnaire
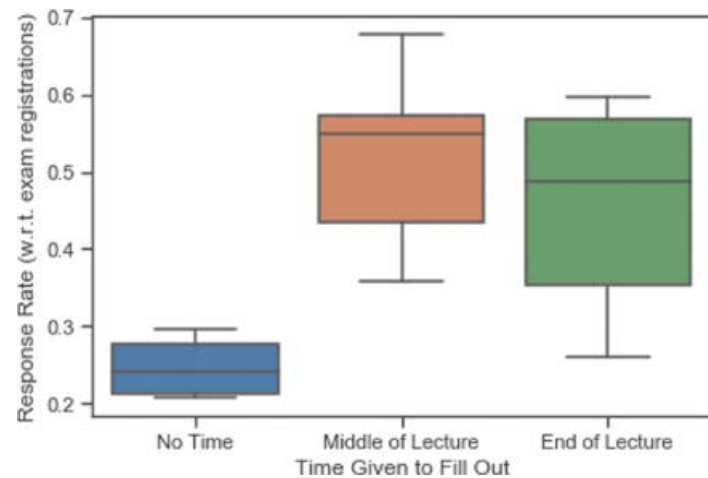
Fig. 7: Response rate with respect to exam registrations of the investigated lectures, grouped by the time provided to answer the questionnaire

If the response rate is measured with respect to lecture registrations, we find that significantly more students fill out the questionnaire when time is given in the middle of the lecture compared to none (p-value 0.014). On average, the response rate was increased by 26 percentage points. With a p-value of 0.068 we have strong reasons to believe that similar conclusions can also be made for time provided at the end of a lecture, however still not being significant at 0.05 level. We can also identify a small trend suggesting that providing time in the middle of the lecture might be better than in the end. These findings presented in table 1 are, however, only small and not statistically significant for the provided sample size.

The same conclusions can also be drawn if exam registrations are taken as a basis for calculating the response rate (Table 2). In this case both time in the middle and at the end significantly increase response rate at significance level 0.05. Thus, our findings are independent of the two used definitions of response rate.

Finally, we can conclude that in order to receive feedback from most students during an online evaluation, each lecturer should provide time around the middle of the lecture for the students to fill out the questionnaires. On average the response rate increases by more than 25 percentage points with a 95% confidence interval between 5 and 48 percentage points.

	Mean Difference of Response Rates	p-value	Interpretation
Middle of Lecture vs. End of Lecture	-0.076	0.536	Not significant
No Time vs. Middle of Lecture	0.260	0.014	Time provided around the middle of the lecture significantly increases response rates compared to not providing time at all
No Time vs. End of Lecture	0.184	0.068	Not significant

Table 1: Results of the case study investigating the effect of provided time slot, based on response rates with respect to lecture registrations

	Mean Difference of Response Rates	p-value	Interpretation
Middle of Lecture vs. End of Lecture	-0.063	0.655	Not significant
No Time vs. Middle of Lecture	0.271	0.013	Time provided around the middle of the lecture significantly increases response rates compared to not providing time at all
No Time vs. End of Lecture	0.208	0.045	Time provided at the end of the lecture significantly increases response rates compared to not providing time at all

Table 2: Results of the case study investigating the effect of provided time slot, based on response rates with respect to exam registrations

2. Impact of the Evaluation Medium

Using the collected data, we assess if online evaluation gives a higher feedback rate than paper based evaluation. The summarized data is shown in table 3. Paper based evaluation was conducted in winter semester 2018/19 and online based evaluation in winter semester 2019/20.

It seems more comparable to use the online evaluation "with time" as comparison. While it is unknown at which times during the lectures paper based evaluation was conducted, it is generally hard to do a paper based evaluation outside the lecture. In this case an online based evaluation increased the mean response rate by 9 percentage points. With p values of 0,20 (w.r.t. lecture registration)

Group	Response Rate w.r.t. Lecture Registrations	Response Rate w.r.t. Exam Registration
Paper based evaluation	35.4%	39.9%
Online based evaluation	39.1%	42.1%
- No time provided	23.0%	24.7%
- With time provided (at the middle or end of lecture)	44.9%	48.4%

Table 3: Results of the case study investigating the impact of the evaluation medium

and 0,26 (w.r.t. exam registration) this difference is not statistically significant.

If, however, the lecturer does not provide a time slot during the lecture, response rates drop by 12 (p value: 0.05) and 15 percentage points (p value: 0.04) respectively. Those findings are statistically significant. However, as argued, this is less meaningful since both provided time and medium are changed.

All in all it seems that an online based evaluation conducted under similar circumstances increases the response rate slightly, but not significantly. The hypothesis of declining number of responses could only be proven if no time was offered to fill out the questionnaire during lectures.

3. Impact of Question Types

In this section, we compare our results that have been conducted at the faculty of chemistry. As mentioned above, we will qualitatively compare two versions of an evaluation questionnaire with the performed Teaching Analysis Poll (TAP) and highlight the most important results: The current form focuses on single choice questions, while the modified form is shortened and focuses on open questions.

Firstly, we compare the replies from the open questions of the two questionnaires. It is noteworthy that the overall amount of answers to open questions has tripled from the current (29%) to the modified questionnaire (79%). Moreover, the respondents used about 2.5 times more words in the modified survey in comparison to the current questionnaire. Not only did the participants write longer and more answers; it should also be noted that the number of presented positive arguments has nearly tripled.

Subsequently, we present the results of the single and multiple-choice questions. Overall, we discover that the results of both questionnaires regarding the single choice questions are relatively similar, even though the current survey has its focus on single choice questions and the modified questionnaire has considerably less questions of this type. For example, a question regarding "The lecturer shows interest in the learning success of his students" has been graded in both questionnaires with an average of 1.6. Furthermore, one question regarding the "Interesting contents" was graded with an average of 1.6 in the modified survey. The questions regarding this context in the current

questionnaire were marked with an average of 1.8 ("Interesting topics"), 1.4 ("Interest of the student"), 2.3 ("suitable contents").

Finally, the results of TAP are compared with the two questionnaires. Altogether, it can be said that answers to the open questions in the modified questionnaire are much more comprehensive than the answers in the current version.

Discussion

General
As each case study was only conducted at one single faculty, the sample size is small. If needed, our findings should be confirmed using a larger sample size and including multiple faculties.

1. Effect of Provided Time Slot
As expected, our study shows that providing time within the lecture to fill out the evaluation form influences the feedback rate positively. This is also underlined by the response rates expected by [1,3,6]: The response rates with no time slot for evaluation provided fall remarkably below the average values of 30-70%, while the response rates with time provided fall in this range.

Although the case study was explained to the lecturers and even though we were in contact with lecturers and students, it is not entirely sure if the time slot for evaluation was provided as described. To increase the results' accuracy, we check the timestamp of every lecture. To give an example, in one lecture, that was assigned to provide time in the middle of the lecture, we observe a large number of timestamps towards the end of a lecture. Asking students who visited this lecture, confirmed that the time was provided towards the end of the lecture and the group is thus changed. Results like this add bias to our study.

2. Impact of Evaluation Medium
The response rates with respect to exam registrations as well as those with respect to lecture registrations lie in the lower range expected in the literature [1,3,6]. Despite this fact, the results seem reasonable considering the high tendency of TUM students to register for exams and lectures without attendance intention.

It is not recorded at which time during the lecture the paper based evaluation was conducted, whereas we collected this data for online

based evaluation. We can thus not perfectly control for this variable when comparing paper based with online evaluation. Since we are not able to find a significant difference between those time points, we assume that our results would only slightly change. Nevertheless, this should be done in further research to receive clearer results. Additionally, data for paper based and online evaluation comes from different years, which is another uncontrolled variable in our study. We could not show a significant difference between the evaluation medium used. This is in line with the results of many other studies [6,12,28,30].

3. Questionnaire Design
Our study suggests that shorter questionnaires lead to more high quality open response feedback (increasing from 29% of the students responding to open feedback questions to 79%), which was also suggested by Sorenson & Reiner [27]. The response rates from the current questionnaire can be classified as at the lower edge of the literatures' estimates [2,22], while those from the modified questionnaire lie on the upper edge.

It should be noted that some questions from the current survey have not been asked in the modified questionnaire. This leads to problems concerning the qualitative comparison of the two questionnaire versions: Negative questions in the current questionnaire, which are not part of the modified version, might not be mentioned by the students' responses in the modified questionnaire. Additionally, the given single choice questions might influence the students' thinking. Another disadvantage of the shortened questionnaire is that the actual number of students supporting an aspect is not known if this aspect appears only in the open questions.

Conclusion
Evaluating lectures is a complex topic, especially if the evaluation is interconnected with a broad quality management system as at TUM. Although having already obtained a well-elaborated evaluation system at TUM, our study suggests that small changes can significantly improve the evaluation system even further:
1. A lecturer should provide time within the lecture for the students to fill out the questionnaire. By doing so, the number of feedback received can be significantly increased by approximately 25 percentage points. Our study suggests that providing time around the middle of the lecture might slightly increase the response rate compared to the end of the lecture. This trend is however not statistically significant.

2. The medium of evaluation (paper vs. online) does not significantly change the response rates, even though more students are reached – including those not attending the lecture regularly.
3. The results of differently designed questionnaires suggest to drastically shorten the single choice questions in favor of more high quality feedback received through open response questions. The shortening of single choice questions does not lead to remarkably decreased feedback interpretation possibilities. Answers to the open questions increased by 150% and contain more positive and longer aspects. Similarly, the comments are much more comprehensive in the modified questionnaire.

To conclude, hypotheses 1 and 3 are confirmed, whereas hypothesis 2 is neither confirmed nor rejected. Further research is needed to strengthen our results: It would be helpful to repeat the conducted studies with more lectures from different faculties. Also different forms of teaching, such as seminars and laboratories, should be added.

Regarding the currently growing online lectures, the results of the case studies investigating the effect of provided time slots and the impact of question types are still valid. Also for online lectures via Zoom or a similar tool, providing time will increase the response rate and open questions will lead to more detailed feedback.

It would also be interesting to assess other methods of increasing the focus on open response questions. Kungliga tekniska högskolan (KTH) in Stockholm for example furnishes every single choice question with a small text field, where a student can comment on the answer he or she has given to the single choice question. This has the advantage that these answers are directly put into context. The effect on feedback quality is unknown as of now, but could be evaluated in a follow up study. More research could investigate the influence of course evaluation after examinations, or publishing regularly the evaluation results of all classes. Both aspects are considered at ETH Zürich and might show a Best-Practice example.

References

1. Adams MJD, Umbach PD (2012) Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. Res High Educ 53: 576–591. https://doi.org/10.1007/s11162-011-9240-5

2. Alhija FN-A, Fresko B (2009) Student evaluation of instruction: What can be learned from students' written comments? Studies in Educational Evaluation 35: 37–44. https://doi.org/10.1016/j.stueduc.2009.01.002

3. Arnold IJM (2009) Do examinations influence student evaluations? International Journal of Educational Research 48: 215–224. https://doi.org/10.1016/j.ijer.2009.10.001

4. CHEN Y, HOSHOWER LB (2003) Student Evaluation of Teaching Effectiveness: An assessment of student perception and motivation. Assessment & Evaluation in Higher Education 28: 71–88. https://doi.org/10.1080/02602930301683

5. Cohen PA (1981) Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. Review of Educational Research 51: 281–309. https://doi.org/10.3102/00346543051003281

6. Dommeyer CJ, Baum P, Hanna RW et al. (2004) Gathering faculty teaching evaluations by in class and online surveys: their effects on response rates and evaluations. Assessment & Evaluation in Higher Education 29: 611–623. https://doi.org/10.1080/02602930410001689171

7. Engels S, Lehner S, Paripovic J et al. Projects in Prospect 2019: EvaluaTUM. In: President of TUM, Prof. Dr. Dr. h.c. mult. Wolfgang A. Herrmann (ed) Research Reports, 2017/II, pp 144–145

8. Feldman KA (1989) The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Res High Educ 30: 583–645. https://doi.org/10.1007/BF00992392

9. Gray M, Bergmann BR (2003) Student Teaching Evaluations: Inaccurate, Demeaning, Misused. Academe 89: 44–46

10. Kersting M, Lang JWB Langfristige Effekte von regelmäßigem Feedback aus studentischen Lehrveranstaltungsevaluationen. Lehrstuhl und Institut für PsychologieUR - http://publications.rwth-aachen.de/record/89645

11. Kreiter CD, Lakshman V (2005) Investigating the use of sampling for maximising the efficiency of student-generated faculty teaching evaluations. Med Educ 39: 171–175. https://doi.org/10.1111/j.1365-2929.2004.02066.x

12. Leung DYP, Kember D (2005) Comparability of Data Gathered from Evaluation Questionnaires on Paper and Through the Internet. Res High Educ 46: 571–591. https://doi.org/10.1007/s11162-005-3365-3

13. Lewis KG (2001) Making Sense of Student Written Comments. New Directions for Teaching and Learning 2001: 25. https://doi.org/10.1002/tl.25

14. L'Hommedieu R, Menges RJ, Brinko KT (1990) Methodological explanations for the modest effects of feedback from student ratings. Journal of Educational Psychology 82: 232–241. https://doi.org/10.1037/0022-0663.82.2.232

15. Lindahl MW, Unger ML (2010) Cruelty in Student Teaching Evaluations. College Teaching 58: 71–76. https://doi.org/10.1080/87567550903253643

16. Loos P, Schäffer B (2001) Das Gruppendiskussionsverfahren: Theoretische Grundlagen und empirische Anwendung. Qualitative Sozialforschung, vol 5. VS Verlag für Sozialwissenschaften, Wiesbaden

17. Marsh HW, Hocevar D (1991) Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. Teaching and Teacher Education 7: 303–314. https://doi.org/10.1016/0742-051X(91)90001-6

18. Marsh HW, Hocevar D (1991) The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. Teaching and Teacher Education 7: 9–18. https://doi.org/10.1016/0742-051X(91)90054-S

19. Marsh HW, Overall JU (1980) Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. Journal of Educational Psychology 72: 468–475. https://doi.org/10.1037/0022-0663.72.4.468

20. Marsh HW, Roche LA (1997) Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. American Psychologist 52: 1187–1197. https://doi.org/10.1037/0003-066X.52.11.1187

21. Morgan DL (2009) Focus groups as qualitative research, 2. ed.,. Qualitative research methods series, vol 16. Sage Publ, Thousand Oaks, Calif.

22. Oliver B, Tucker B, Pegden (2007) An investigation into studentcomment behaviors: Who comments, what do they say, and do anonymous st udentsbehave badly? In: Coates H (ed) Developing generalisable measures of knowledge and skill outcomes in higher education. Australian Universities Quality Agency

23. Penny AR, Coe R (2004) Effectiveness of Consultation on Student Ratings Feedback: A Meta-Analysis. Review of Educational Research 74: 215–253. https://doi.org/10.3102/00346543074002215

24. Rantanen P (2013) The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. Assessment & Evaluation in Higher Education 38: 224–239. https://doi.org/10.1080/02602938.2011.625471

25. Rindermann H, Schofield N (2001) Generalizability of Multidimensional Student Ratings of University Instruction Across Courses and Teachers. Research in Higher Education 42: 377–399. https://doi.org/10.1023/A:1011050724796

26. Schneider M, Mustafić M (2015) Gute Hochschullehre: Eine evidenzbasierte Orientierungshilfe. Springer Berlin Heidelberg, Berlin, Heidelberg

27. Sörenson DL, Reiner C (2003) Charting the Uncharted Seas of Online Student Ratings of Instruction. New Directions for Teaching and Learning 96: 1–24

28. Stowell JR, Addison WE, Smith JL (2012) Comparison of online and classroom-based student evaluations of instruction. Assessment & Evaluation in Higher Education 37: 465–473. https://doi.org/10.1080/02602938.2010.545869

29. TUM Center for Study and Teaching, Studium und Lehre – Qualitätsmanagement QM-Handbuch zum Studiengangslebenszyklus: Prozesse, Wegweiser und Vorlagen. http://www.lehren.tum.de/?id=422#c6260. Accessed 17 Jul 2020

30. Venette S, Sellnow D, McIntyre K (2010) Charting new territory: assessing the online frontier of student ratings of instruction. Assessment & Evaluation in Higher Education 35: 97–111. https://doi.org/10.1080/02602930802618336

31. Zimmaro DM, Gaede CS, Heikes EJ et al. Running Head: Study of Written Course Evaluation Comments

Self Reflection

From the very beginning, our team had a clear vision: All of us wanted to improve the evaluation system at the Technical University of Munich (TUM). Therefore, we could focus directly on the project and had not to search for a concrete goal. Due to our small team size of five members this was very efficient.

However, we had difficulties in the beginning. Two of us were in Zurich for the first 9 months. We could thus not meet in person, which led to less personal contact. Also our supervisor and tutors were not named before January. However, until the end of January, we already prepared our team structure, which we had to adjust later on.

After the kickoff event, our project plan was to do literature review and get in contact with TUM administration. In the following we wanted to structure our ideas, derive a hypothesis and plan our methodology. These methods should have been discussed with different faculties and then be implemented. The final step was to evaluate our data gained and write the report.

The literature research turned out to be a difficult start. As none of us had insights and knowledge in the field of evaluation, the correct keywords to find useful references were missing. Only looking up for the buzzword "evaluation" results in literature about evaluation of industrial processes and economical services. None of us knew how we could get deeper into the material. As we also did not have a supervisor at this part of the project, we kept our literature review very superficial. This might have been a problem that slowed us down.

While communicating with different people from TUM we learned many lessons. However, these were not only things, which helped us with our project. In many cases, we learned about communica-tion and miscommunication. We learned about difficulties contacting people via email and about problems meeting people in person. We figured out that communication could make people think that there are problems, while there are actually none. All these issues made us learn a lot, but threw us also back regarding our schedule. Due to conflicts resulting from (mis)communication, our project partners had not yet been available. Therefore, we developed ideas without partners first and searched for partners later on. This plan had its advantages but also disadvantages. As we already had clear concepts, we were able to present our prospective partners a clear view over possible implementations of our ideas. However, we had to develop more concepts, to be adaptive for our partners. This also led to the point that not necessarily our favourite models were implemented.

Before we were able to present the concepts to prospective part-ners, we had to develop them. Somehow harder than gaining ideas, was to specify the most useful and reasonable ones. Therefore, we created a document and rated all ideas by different factors. This was on the one hand a fast and efficient possibility. On the other hand, it was a partly random procedure. To overcome this, we planned some expert interviews to rate our ideas. Therefore, we had to find different people who are involved in teaching (e.g. lecturers, quality managers, students). As we were already quite late in our process, we skipped this part, which we should have done earlier.

The implementation phase began during the summer break. This worked out better than expected, even if there were usually three of the five of us on holiday. The issue was more after the summer holidays: It was hard to motivate ourselves and our structures did not help us. Unfortunately, our supervisor was not available for

three months. In this time, our tutor helped us a lot to restructure ourselves and we developed again at a faster pace. One change we made was to name two members responsible for each case study with the Faculty of Chemistry and the Munich School of Engineering. They were each responsible for the overview and coordination of the team members. This was a well-timed change at the right moment. However, also not being too strict in earlier stages helped us that all worked together.

In November, the main part of our implementations started. Even though some minor problems occurred, our clear target helped us to overcome these issues. In this phase of our project, we had an intense exchange with our tutor to get everything done as good as possible. However, we noticed that our motivation decreased for our second implementation. Beside the fact that we already had done one project part and were less motivated to overcome more unexpected difficulties, also the exam period made team members less involved.

Looking at the whole project, it was difficult for us to find established concepts to measure the outcome of our experiments. To find the final chosen measurement, we had two main steps. The first one was the statistical consulting offered by the TUM. This helped us a lot to interpret the quantitative results correctly. On the other hand, we also wanted to make a qualitative analysis. Therefore, we searched for expert knowledge and got in touch with Mrs. Spiekermann and ProLehre, a TUM institution. They gave us an idea how to get a benchmark and also trained us for that.
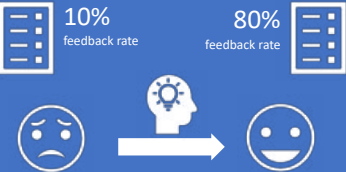
As during the whole project phase, always at least one of our team members was abroad, the TUM: Junge Akademie supported us by a weekend trip to our members abroad. We met in Zurich and developed our project ideas. A second meeting was planned in Stockholm, but got cancelled due to the corona pandemic. Beside the TUM: Junge Akademie weekend meetings, our only team meeting in person over the whole 20 months period, remained the weekend in Zurich. ∎

Process Description

Kickoff:

10% feedback rate

80% feedback rate

Organisation, Planning
(Poster 1)

Communication

Survey
(Poster 2)

Ideas & Concepts
(Poster 2)

Communication
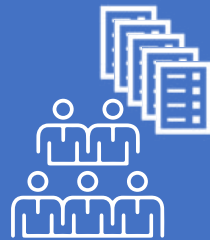
EvaluaTUM

OUR IDEA AT A GLANCE

As our name "EvaluaTUM" indicates, we want to analyse the existing evaluation system at TUM. From our experiences, there are various approaches and therefore differently helpful results at different faculties while the field of lecture evaluation is complex and many perspectives need to be considered.

GOAL

Our overall goal is to improve teaching at TUM until June 2020 by taking into account two approaches:

1. On the one hand we want to increase the quality of communication between students and professors, measured by an outcome evaluation.
2. On the other hand, we want to improve the response rate of evaluations.

PROJECT STRUCTURE PLAN
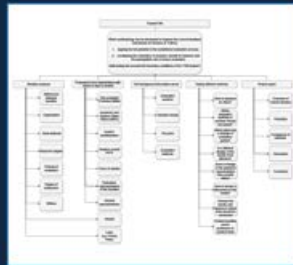
RESEARCH QUESTION

Which methodology can be developed to improve the current feedback instruments for lectures at TUM by

1. tapping the full potential of the established evaluation process,
2. considering the motivation of students, benefit for lecturers and the participation rate in lecture evaluation

while taking into account the boundary conditions of the TUM system?

SUB QUESTIONS

1. Where is potential for improvement in lecture evaluation at the TUM from a students' and lecturers' perspective?
2. Which methods can be developed using the identified potentials?
3. How do different evaluation methods influence the satisfaction of the lecturers as well as the motivation of the students?
4. How can we find a balance between individual teaching assessment and the constraint of comparability?

TIME SCHEDULE

November 2018:
— Begin of project
November 2018 – June 2019:
— Situation Analysis
May – July 2019:
— Development of methods to test
May – October 2019:
— Organization of experiments (faculties, lectures, method refinement and adjustment)
October 2019 – March 2020:
— Scientific experiment
February – April 2020:
— Evaluation of results
June 2020:
— Final Symposium

MAY 2019

MEMBERS Stefan Engels, Stefan Lehner, Judith Paripovic, Pascal Resch, Paul Sieber
TUTORS Konstantin Riedl, Dennis Goldner
MENTOR Prof. Dr. Annette Noschka-Roos

inspired by TUM: Junge Akademie

EvaluaTUM

THE IDEA AT A GLANCE

As our name "EvaluaTUM" indicates, we want to analyse the existing evaluation system at TUM. From our experiences, there are various approaches and therefore differently helpful results at different faculties while the field of lecture evaluation is complex and many perspectives need to be considered.

GOAL

Our overall goal is to improve teaching at TUM until June 2020 by taking into account two approaches:

1. On the one hand we want to increase the quality of communication between students and professors, measured by an outcome evaluation.
2. On the other hand, we want to improve the response rate of evaluations.

RESEARCH QUESTION

Taking into account the boundary conditions of the TUM system, what sort of methodology can be developed to improve the current feedback instruments for teaching at TUM that will

1. tap the full potential of the evaluation process and
2. address the motivation of students, the benefits for lecturers, and the participation rate in lecture evaluation?

WHAT HAPPENED SO FAR?

In the last months, we conducted a comprehensive analysis of the current situation at TUM concerning the course evaluation. At the moment we are in the process of finalizing this research by analyzing and existing questionnaires and the "teacher survey" which recently took place at TUM.

Our goal is to get to know the evaluation process from different perspectives, to establish the right toeholds. Thus, during the winter semester 2019/20, we are going to implement and compare different methods for lecture evaluation with respect to satisfiability, quality of feedback and response rate. We will decide which methods to test in the next weeks after our situational analysis is completed.

WHERE AND HOW SHOULD YOU ADAPT YOUR APPROACH?

Contact us: evaluatum@ja.tum.de

JUNE 2019

MEMBERS Stefan Engels, Stefan Lehner, Judith Paripovic, Pascal Resch, Paul Sieber
TUTORS Konstantin Riedl, Dennis Goldner
MENTOR Prof. Dr. Annette Noschka-Roos

inspired by TUM: Junge Akademie