Position statement of the GMA committee “teaching evaluation”

Abstract

The evaluation of teaching can be an essential driver for curriculum development. Instruments for teaching evaluation are not only used for the purpose of quality assurance but also in the context of medical education research. Therefore, they must meet the common requirements for reliability and validity. This position paper from the GMA Teaching Evaluation Committee discusses strategic and methodological aspects of evaluation in the context of undergraduate medical education and related courses; and formulates recommendations for the further development of evaluation. First, a four-step approach to the design and implementation of evaluations is presented, then methodological and practical aspects are discussed in more detail. The focus here is on target and confounding variables, survey instruments as well as aspects of implementation and data protection. Finally, possible consequences from evaluation data for the four dimensions of teaching quality (structural and procedural aspects, teachers and outcomes) are discussed.

Keywords: evaluation, teaching, teaching quality, reliability, validity, medicine, health sciences

Nicolas Haverkamp[1]
Janina Barth[2]
Dennis Schmidt[3]
Uta Dahmen[4]
Oliver Keis[5]
Tobias Raupach[6]

1 University of Bonn, Medical Faculty, Office of the Dean of Studies, Bonn, Germany

2 University of Lübeck, Department of Human Medicine Studies and Teaching, Lübeck, Germany

3 University Medical Center Göttingen, Office of the Dean of Studies, Göttingen, Germany

4 University of Jena, Medical Faculty, Clinic for General, Visceral and Vascular Surgery, Jena, Germany

5 University of Ulm, Medical Faculty, Office of the Dean of Studies, Ulm, Germany

6 University of Bonn, Medical Faculty, Institute for Medical Didactics, Bonn, Germany

Introduction

The evaluation of teaching is the final element of the Kern cycle and should be a key driver for curriculum development [1]. Depending on the data collection instrument and the people surveyed, evaluations can collect information about structural and procedural aspects of teaching, about the behavior of individual teachers and ultimately about student learning outcomes [2]. On the one hand, teaching evaluation is suitable as an instrument for quality assurance and improvement [3], and on the other hand, evaluation instruments are also used in medical education research. In both cases, the procedures used should meet high standards of scientific work such as sufficient reliability and validity of the data collection tools, since numerous potential confounding factors must be taken into account when collecting and interpreting student evaluation data and sometimes far-reaching consequences are drawn from evaluation results.

This position paper from the GMA Teaching Evaluation Committee discusses strategic and methodological aspects of evaluation in the context of medical studies and related study programs and formulates specific recommendations for evaluation [4] in addition to general standards for evaluation that have already been formulated. The position paper is intended to serve as a starting point for medical schools to further develop their own evaluation processes; it is also intended to contribute to the creation of a more general evaluation concept for undergraduate medical education and related courses.

Background

The discussions of the GMA Teaching Evaluation Committee crystallized a desire for developing questionnaires that would facilitate a comprehensive evaluation of teaching across courses and locations. This would allow comparisons between medical schools – also in view of the revision of the National Competence-Based Catalog of Learning Objectives for medicine/dentistry and the upcoming reform of the Medical Licensure Act. In connection with the digitisation of teaching since the summer term of 2020, various questionnaires have already been developed by members of the committee in a corresponding pilot project, which are aimed at both teachers and students and look at different levels of teaching organisation (meso and micro levels). These forms were made available to all medical schools.

A general data collection instrument will not be able to cover all dimensions of teaching quality in the same way, so prioritizing content seems sensible here. The design of an instrument that can be used universally and does not require additional hardware and software at different sites remains a challenge. There is also a need for new survey instruments that take current developments (e.g. emphasis on communication and interprofessionalism) and special teaching settings (e.g. teaching practical skills, e.g. in skills labs) into account.

Basic considerations

The Committee recommends the following four-step approach:

1. **Defining the objective(s) of the evaluation:** Before an evaluation is carried out, the aim of the evaluation must be clearly stated and agreed upon. This also includes defining a construct to be measured as well as considerations for the exclusion of possible confounding variables. If possible, it should be determined in advance what consequences are to be drawn from the results.
2. **Selection of data collection instrument(s):** The central step in the evaluation process is the selection of one or more suitable instruments to achieve the stated goal. This requires a decision on whether to choose an existing survey instrument with known psychometric properties or whether a new instrument must be developed. For specific purposes it may be useful to use non-validated survey forms; however, in general and from a scientific perspective, it is recommended to use published and validated survey instruments.
3. **Data collection:** In principle, it must be clarified when or during which period the data should be collected and whether data collection should be paper-based or online. In the latter case, various platforms are available, some of which can automatically send notifications and reminders to the target audience.
4. **Presentation of the results:** An appropriate approach to data analysis and presentation should be taken in order to facilitate a meaningful interpretation of the data. Analyses that go beyond descriptive evaluations should only be carried out in justified cases.

Depending on the objectives, the evaluation process should have the support of the medical school committees. Responsibilities, workflows and the implementation of consequences should be communicated transparently throughout.

Methodological and practical aspects

In the following, individual aspects of the evaluation methodology will be discussed – in some cases with reference to the four steps mentioned above.

1. Defining the objective(s) of the Evaluation

The first step in the evaluation process is determining the evaluation goal. Depending on this goal, suitable survey instruments will be selected. In general, the aim of evaluation in most cases is to measure the quality of teaching and to strive for improvements on this basis. For this purpose, it is necessary to define the construct "good teaching" for the context at hand. For example, in a survey of around 800 clinical lecturers at German medical schools, "good teaching" was characterized in particular by its leading to sustainable learning outcomes among students and increased interest in the content taught [5].

Target variable(s)

Two definitions of good teaching – the Stanford criteria and the four dimensions according to Gibson – will be portrayed in greater detail here because they are widely known and easy to operationalise. In addition, data collected based on these definitions can serve as a solid starting point for ongoing curriculum development. There are of course other classifications and systems for evaluating "good teaching" as well, all of which allude to similar dimensions [6], [7].

The Stanford criteria [8], originally developed as part of a faculty development program, are:

1. Establishing a positive learning climate,
2. Control of the teaching session,
3. Communicating goals,
4. Promoting understanding and retention,
5. Evaluation,
6. Feedback and
7. Promoting self-directed learning.

According to Gibson [2], teaching quality is a multidimensional construct. In advance of the evaluation, it must be

clarified which dimension(s) of good teaching are to be assessed:

- The structural dimension includes the overall conditions under which teaching takes place, including the teaching facilities, the basic structure of a course or the existence of a catalogue of learning objectives.
- The procedural dimension refers to the teaching sessions themselves. Quality indicators commonly used here include, for example, the punctuality of teaching staff and students, the type of interaction and the use of different teaching methods [9], [10]. A term often used in this context is the "learning climate" [11].
- The personal dimension refers to the teaching staff. In most cases, teachers have little opportunity to influence structural aspects of teaching. However, they can certainly shape teaching processes. Due to the fact that teachers play an important role in most curricula, specific survey instruments have been developed to assess the performance of individual teachers [12], [13], [14], [15]. These can also be used as part of assessments relevant to career development.
- The outcomes dimension of teaching quality refers to the measurable result of teaching. This is often also referred to as the students' learning success (although other definitions are also possible). Examinations can be used to measure learning success; however, these only reflect the level of performance at the end of a teaching intervention and do not say anything about the actual increase of knowledge, skills and professionalism during the intervention. As an alternative to determining learning success, repeated self-assessments can also be used, the reliability and validity of which have been examined in some studies [16], [17], [18].

Students are often asked – without reference to any of the aforementioned dimensions – to provide a general appraisal of teaching, for example by awarding school grades [19].
Such a global rating merely corresponds to satisfaction. Due to the uniformity of the school grade system, the results suggest comparability between courses. However, satisfaction ratings based solely on school grades are particularly difficult to interpret because the construct of satisfaction is very individual. In addition, not all teaching formats that students are satisfied with are effective. Conversely, not all teaching formats that have been proven to be effective lead to increased satisfaction among students.

Confounding variables

Any impact of construct-irrelevant factors poses a threat to the validity of target variable measurements. The risk of confounding is particularly high if the construct to be measured has not been determined precisely enough – as a consequence, it can be difficult to distinguish between valid influences and confounding influences on measurements.

A literature review from 2015 identified numerous factors that can impact student global assessments (e.g. school grade ratings) of courses [20]. While some of these factors (e.g. course organization, communication and feedback for students) are easily compatible with the construct of good teaching, others can clearly be described as confounding variables (e.g. gender of students, interest of students in the content taught, time of data collection). Two aspects should be particularly emphasised at this point:

- Evaluation format: In one study, online data collection was associated with a lower response rate than paper-based surveys [21], [22]. At the same time, the ratings for online data collection were more positive than for paper-based data collection.
- Voluntary participation in evaluation: If participation in the evaluation is voluntary, high-performing students may be more likely to take part; this can lead to the evaluations being more positive than if all students took part. Conversely, an obligation to evaluate (if at all possible from a legal perspective) can lead to some students not taking the exercise serious enough, thus increasing construct-irrelevant variance in the data.

Although it seems hardly possible to take all confounding variables into account or adjust for them when analysing data (because due to data protection, no personal information such as the gender of participants is collected), the people who carry out evaluations and interpret their results should be aware of these effects and, if necessary, account for them when planning an evaluation.

2. Selection of data collection instrument(s)

The second step in the evaluation process is the selection of appropriate evaluation tool(s). A typical target audience for teaching evaluations are students. Alternatively or in addition to this, graduates or patients can also be asked to evaluate specific aspects of the teaching. Lecturers can use self-assessments (under supervision, e.g. in faculty development programs) or peer ratings to obtain feedback on the quality of their own teaching. Finally, other existing data such as progress logs or aggregated examination results can also be used to evaluate teaching.
Depending on the data source and the evaluation goals, different data collection formats and instruments are used:
According to a 2021 survey, the most common format used by 97% of medical schools [23] is the questionnaire, either to be completed on paper or online. A survey among German medical schools showed that online questionnaires are used at almost all locations. Half of the schools also use paper-based questionnaires [23], [24]. There are numerous general questionnaires available for evaluating academic teaching; however, aspects specific to medical education such as bedside teaching are not always adequately taken into account. For this reason,

specific instruments for assessment in medical and related courses have been developed and scientifically validated.
Many of the questionnaires used by individual medical schools aim to evaluate the quality of teaching locally. Due to the specifics of some sites which are reflected in the evaluation forms, such forms normally cannot be used across medical schools. In contrast, published questionnaires with known psychometric characteristics offer the advantage of being largely location-independent; however, according to the faculties' own information, these are currently only used at 7% of sites [23]. Combinations of validated and non-validated forms are used by 30% of medical schools [23]. While this is possible, this should be stated clearly.
Although questionnaires can contain free text fields, in most cases the questionnaires are used to collect quantitative data. In contrast, individual interviews or focus groups are more suitable for generating qualitative data. According to the medical schools, focus groups are used in 21% and interviews in 15% of medical courses in Germany [24]. There are also course debriefings, which can be more or less structured and primarily serve to discuss specific problems and solutions.
When constructing questionnaires with scaled items, research findings on questionnaire construction must, among other things, be taken into account [25], [26].

3. Data collection

Aspects relevant to the implementation of evaluations include the type of data collection (online, paper-based, personal or telephone interviews), curricular integration (mandatory vs. voluntary participation) and the time of data collection. Survey instruments and mandatory vs. voluntary participation have already been discussed. This section addresses the timing of evaluations. Evaluation often takes place at the end of a lecture, a course/module or a term/ study section. While teaching staff get an ideal overview of the quality of teaching in this way, this timing can have a negative impact on the motivation of students because they do not experience the consequences of the evaluation themselves. Furthermore, data collection at the end of term can have a negative impact on criterion validity due to the delay between the survey and the subject of the evaluation which is often too long. At some sites, evaluations are therefore held in the middle of a course in order to be able to make adjustments in the current term. However, too much data collection in quick succession can lead to evaluation fatigue among students, which can have a negative impact on response rates. Methods were therefore developed that make it possible to draw generalisable conclusions from feedback from a small part of the student cohort. This method was initially only examined for satisfaction ratings [27], [28]. However, a recent study shows that a reliable assessment of student learning outcome is possible even if only around 20-30% of the students in a cohort take part in the evaluation [29].

A special feature of medical education is that – especially in courses with a clinical focus – there is usually not just one single teacher instructing the entire course. Instead, a large number of teaching staff are involved in supervising smaller groups of students. This circumstance must be taken into account when evaluating individual teaching performance, e.g. by allowing students to evaluate the teachers they have been assigned to individually and shortly after the respective session.
In order to avoid confounding by dissatisfaction with individual examination performance, an evaluation of teaching should be carried out before taking the examination – although this is often difficult to implement in reality. If the examination itself is to be subjected to an evaluation by students, separate data collection must be planned for this.
The workflow for carrying out an evaluation within a medical school, as well as the consequences of its results in the broadest sense, are also part of the implementation; due to the great importance of these aspects, separate chapters are dedicated to them.

4. Presentation of the results

When choosing a suitable presentation format for evaluation results, various aspects should be kept in mind:
Addressees should be able to discern the sample size, the response rate, the data distribution and any floor and ceiling effects from the presentation.
Inferential statistical analyses should only be carried out if there is a question justifying their use. In this case the procedure should be named.
Outside scientific studies, it should be clearly stated that all analyses are exploratory.

Workflow: Data protection and responsibilities

Ensuring teaching quality is a central task of the dean of studies. The office of the dean of studies should therefore have the personal and technical expertise to carry out evaluations professionally and communicate the results within the medical school. On the one hand, this includes technical aspects of data collection and processing but on the other hand and above all, the competence to describe evaluation objectives and to select congruent instruments based on knowledge of the current literature. Decentralized activities can be useful (particularly for evaluating teaching innovations in defined areas). However, data protection is of particular importance in this context, for example if the performance of individual teachers is to be evaluated. If – as is very likely in this case – there is a need to refer the matter to the respective data protection commission, the relevant deadlines must be observed. Responsibility for handling evaluation data lies either with the office of the Dean (of studies) or with a person named in the respective evaluation regulations. Conversely, it must also be ensured that the people who take part in an evaluation remain anonymous. In

order to collect personal data in an evaluation project, the consent of those involved may be necessary.

Consequences from evaluation results

An overarching goal of teaching evaluation is to identify strengths and weaknesses and – based on this analysis – to continuously improve teaching. The consequences of the evaluation results must be based on the previously defined goals. Accordingly, they can refer to teaching-related structures and processes, teaching staff and/or the teaching outcomes.

- Consequences for the structural dimension of teaching: If evaluation reveals evidence of sub-optimal teaching conditions (rooms, accessibility of digital resources, etc.), improvements must be pursued. The structural conditions of teaching also indirectly include financial resources. However, the derivation of financial consequences of evaluations is heavily debated [18] and requires that a quality-assured evaluation instrument is used and that there is comparability between different courses.
- Consequences for the procedural dimension of teaching: Depending on which optimisation needs arise from the data, measures could be taken with the aim of improving the learning climate, optimising the implementation of instructional formats or supporting learning processes in other ways (e.g. digitally).
- Consequences for the personal dimension of teaching (teaching staff): At many medical schools (82%), feedback discussions are held with teaching staff based on evaluation data [24]. In the context of a personal evaluation, the data can result in direct individual consequences such as the referral to faculty development courses. If teaching is to be given adequate weight in academic careers, promotion to a reader position can be made dependent on the results of multiple teaching evaluations by various actors (e.g. students, experts in medical education, colleagues as critical friends).
- Consequences with regard to teaching outcomes: If the evaluation shows that the desired learning objectives are not being achieved, this should have comprehensive consequences for the constructive alignment within the relevant course (i.e. are the learning objectives appropriate/are the teaching methods adequate/do the exams represent the content in an appropriate format etc.).

Recommendations

- The assessment of teaching quality should be based on reliable data. The Committee therefore recommends that when designing and implementing teaching-related evaluations, particular emphasis should be placed on pairing the desired goals with the right data collection instruments.
- In order to enable cross-location and cross-disciplinary comparisons, validated evaluation instruments should be used whenever possible. Published evaluation instruments that take into account specific aspects of medical education and related courses (e.g. patient-centered teaching formats) are available and should be used accordingly [13], [15].

Note

The position paper was acepted by the GMA executive board at 24-01-2024.

Authors' ORCIDS

- Nicolas Haverkamp: [0000-0001-7814-8425]
- Janina Barth: [0000-0003-1905-1257]
- Uta Dahmen: [0000-0003-3483-3388]
- Tobias Raupach: [0000-0003-2555-8097]

Competing interests

The authors declare that they have no competing interests.

References

1. Kern DE, Thomas PA, Howard DM, Bass EB. Curriculum Development For Medical Education - A Six Step Approach. Baltimore/London: The Johns Hopkins University Press; 1998.
2. Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP. Enhancing Evaluation in an Undergraduate Medical Education Program. Acad Med. 2008;83(8):787-793. DOI: 10.1097/ACM.1090b
3. Berk RA. Top five flashpoints in the assessment of teaching effectiveness. Med Teach. 2013;35(1):15-26. DOI: 10.3109/0142159X.2012.732247
4. Gesellschaft für Evaluation (DeGEval). Standards für Evaluation. Mainz: DeGEval – Gesellschaft für Evaluation e.V.; 2016. Zugänglich unter/available from: https://www.degeval.org/degeval-standards/standards-fuer-evaluation/
5. Schiekirka-Schwake S, Anders S, von Steinbuchel N, Becker JC, Raupach T. Facilitators of high-quality teaching in medical school: findings from a nation-wide survey among clinical teachers. BMC Med Educ. 2017;17(1):178. DOI: 10.1186/s12909-017-1000-6
6. Kirkpatrick DL. Evaluating training programs: The four levels. San Francisco: Berrett-Koehler; 1994.
7. Rindermann H. Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts. Landau: Empirische Pädagogik e.V.; 2001.
8. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med. 1998;73(6):688-695. DOI: 10.1097/00001888-199806000-00016

9. De la Fuente J, Sander P, Fernando J, Pichardo-Martinez MC. Validation Study of the Scale for Assessment of the Teaching-Learning Process, Student Version ATLP-S. Electron J Res Educ Psychol. 2010;8(2):815-840. DOI: 10.25115/ejrep.v8i21.1397
10. Staufenbiel T. Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. Diagnostica. 2000;46(4):169-181. DOI: 10.1026//0012-1924.46.4.169
11. Roff S, McAleer S, Harden RM, Al-Qahtani M, Ahmed AU, Deza H, Groenen G, Primparyon P. Development and validation of the Dundee Ready Education Environment Measure (DREEM). Med Teach. 1997;19(4):295-299. DOI: 10.3109/01421599709034208
12. Iblher P, Zupanic M, Hartel C, Heinze H, Schmucker P, Fischer MR. The Questionnaire "SFDP26-German": a reliable tool for evaluation of clinical teaching? GMS Z Med Ausbild. 2011;28(2):Doc30. DOI: 10.3205/zma000742
13. Dreiling K, Montano D, Poinstingl H, Müller T, Schiekirka-Schwake S, Anders S, von Steinbüchel N, Raupach T. Evaluation in undergraduate medical education: Conceptualizing and validating a novel questionnaire for assessing the quality of bedside teaching. Med Teach. 2017;39(8):820-827. DOI: 10.1080/0142159X.2017.1324136
14. Zuberi RW, Bordage G, Norman GR. Validation of the SETOC instrument -- Student evaluation of teaching in outpatient clinics. Adv Health Sci Educ Theory Pract. 2007;12(1):55-69. DOI: 10.1007/s10459-005-2328-y
15. Müller T, Montano D, Poinstingl H, Dreiling K, Schiekirka-Schwake S, Anders S, Raupach T, von Steinbüchel N. Evaluation of large-group lectures in medicine - development of the SETMED-L (Student Evaluation of Teaching in MEDical Lectures) questionnaire. BMC Med Educ. 2017;17(1):137. DOI: 10.1186/s12909-017-0970-8
16. Schiekirka S, Reinhardt D, Beissbarth T, Anders S, Pukrop T, Raupach T. Estimating learning outcomes from pre- and posttest student self-assessments: a longitudinal study. Acad Med. 2013;88(3):369-375. DOI: 10.1097/ACM.0b013e318280a6f6
17. Raupach T, Münscher C, Beissbarth T, Burckhardt G, Pukrop T. Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness. Med Teach. 2011;33(8):e446-453. DOI: 10.3109/0142159X.2011.586751
18. Schiekirka S, Anders S, Raupach T. Assessment of two different types of bias affecting the results of outcome-based evaluation in undergraduate medical education. BMC Med Educ. 2014;14:149. DOI: 10.1186/1472-6920-14-149
19. Schiekirka S, Feufel MA, Herrmann-Lingen C, Raupach T. Evaluation in medical education: A topical review of target parameters, data collection tools and confounding factors. Ger Med Sci. 2015;13:Doc15. DOI: 10.3205/000219
20. Schiekirka S, Raupach T. A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. BMC Med Educ. 2015;15:30. DOI: 10.1186/s12909-015-0311-8
21. Paolo AM, Bonaminio GA, Gibson C, Patridge T, Kallail K. Response rate comparisons of e-mail- and mail-distributed student evaluations. Teach Learn Med. 2000;12(2):81-84. DOI: 10.1207/S15328015TLM1202_4
22. Adams MJ, Umbach PD. Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. Res High Educ. 2012;53(5):576-591. DOI: 10.1007/s11162-011-9240-5
23. Giesler M, Kunz K. „Closing the Gap“: Ergebnisse einer Bestandsaufnahme zu Qualitätssicherungsmaßnahmen der Lehre an medizinischen Fakultäten ["Closing the Gap": Results of a survey assessing quality assurance in medical education]. Z Evid Fortbild Qual Gesundhwes. 2021;164:51-60. DOI: 10.1016/j.zefq.2021.05.008
24. Schiekirka-Schwake S, Barth J, Pfeilschifter J, Hickel R, Raupach T, Herrmann-Lingen C. National survey of evaluation practices and performance-guided resource allocation at German medical schools. Ger Med Sci. 2019;17:Doc04. DOI: 10.3205/000270
25. Albanese M, Prucha C, Barnet JH, Gjerde CL. The effect of right or left placement of the positive response on Likert-type scales used by medical students for rating instruction. Acad Med. 1997;72(7):627-630. DOI: 10.1097/00001888-199707000-00015
26. Albanese M, Prucha C, Barnet JH. Labeling each response option and the direction of the positive options impacts student course ratings. Acad Med. 1997;72(10 Suppl 1):S4-6. DOI: 10.1097/00001888-199710001-00002
27. Cohen-Schotanus J, Schonrock-Adema J, Schmidt HG. Quality of courses evaluated by 'predictions' rather than opinions: Fewer respondents needed for similar results. Med Teach. 2010;32(10):851-856. DOI: 10.3109/01421591003697465
28. Schönrock-Adema J, Lubarsky S, Chalk C, Steinert Y, Cohen-Schotanus J. 'What would my classmates say?' An international study of the prediction-based method of course evaluation. Med Educ. 2013;47(5):453-462. DOI: 10.1111/medu.12126
29. Grebener BL, Barth J, Anders S, Beissbarth T, Raupach T. A prediction-based method to estimate student learning outcome: Impact of response rate and gender differences on evaluation results. Med Teach. 2021;43(5):524-530. DOI: 10.1080/0142159X.2020.1867714

Corresponding author:
Prof. Dr. Tobias Raupach, MME
University of Bonn, Medical Faculty, Institute for Medical Didactics, Venusberg - Campus 1, D-53127 Bonn, Germany
tobias.raupach@ukbonn.de

Please cite as
Haverkamp N, Barth J, Schmidt D, Dahmen U, Keis O, Raupach T. Position statement of the GMA committee “teaching evaluation”. GMS J Med Educ. 2024;41(2):Doc19.
DOI: 10.3205/zma001674, URN: urn:nbn:de:0183-zma0016745

This article is freely available from
https://doi.org/10.3205/zma001674

Received: *2023-03-16*
Revised: *2023-11-29*
Accepted: *2024-02-01*
Published: *2024-04-15*

Positionspapier GMA-Ausschuss „Lehrevaluation“

Zusammenfassung

Die Evaluation der Lehre kann ein wesentlicher Motor für die Curriculumsentwicklung sein. Instrumente zur Lehrevaluation werden aber nicht nur zum Zweck der Qualitätssicherung, sondern auch im Kontext der medizindidaktischen Forschung eingesetzt. Entsprechend müssen Sie den gängigen Anforderungen an Reliabilität und Validität genügen. In diesem Positionspapier des GMA-Ausschusses Lehrevaluation werden strategische und methodische Aspekte der Evaluation im Kontext des Medizinstudiums und verwandter Studiengänge erörtert und Empfehlungen für die Weiterentwicklung der Evaluation formuliert. Zunächst wird ein vierschrittiges Vorgehen für die Konzeption und Durchführung von Evaluationen vorgestellt, danach wird vertieft auf methodische und praktische Aspekte eingegangen. Schwerpunkte hierbei sind Ziel- und Störvariablen, Erhebungsinstrumente sowie Aspekte der Implementierung und des Datenschutzes. Schließlich werden mögliche Konsequenzen aus Evaluationsdaten für die vier Dimensionen der Lehrqualität (Strukturen, Prozesse, Personen und Ergebnisse) diskutiert.

Schlüsselwörter: Evaluation, Lehre, Lehrqualität, Reliabilität, Validität, Medizin, Gesundheitswissenschaften

Nicolas Haverkamp[1]
Janina Barth[2]
Dennis Schmidt[3]
Uta Dahmen[4]
Oliver Keis[5]
Tobias Raupach[6]

1 Universität Bonn, Medizinische Fakultät, Studiendekanat, Bonn, Deutschland

2 Universität zu Lübeck, Referat Studium und Lehre Humanmedizin, Lübeck, Deutschland

3 Universitätsmedizin Göttingen, Studiendekanat, Göttingen, Deutschland

4 Universität Jena, Medizinische Fakultät, Klinik für Allgemein-, Viszeral- und Gefäßchirurgie, Jena, Deutschland

5 Universität Ulm, Medizinische Fakultät, Studiendekanat, Ulm, Deutschland

6 Universität Bonn, Medizinische Fakultät, Institut für Medizindidaktik, Bonn, Deutschland

Einführung

Die Evaluation der Lehre bildet das letzte Element im Kern-Zyklus und sollte ein wesentlicher Motor für die Curriculumsentwicklung sein [1]. Je nach Datenerhebungsinstrument und befragten Personen können in Evaluationen Informationen über strukturelle und prozedurale Aspekte der Lehre, über das Verhalten einzelner Lehrender und schließlich auch über den studentischen Lernerfolg erhoben werden [2]. Die Lehrevaluation eignet sich zum einen als Instrument der Qualitätssicherung und -verbesserung [3], zum anderen kommen auch in der medizindidaktischen Forschung Evaluationsinstrumente zum Einsatz. In beiden Fällen sollten die genutzten Verfahren hohen Ansprüchen an wissenschaftliches Arbeiten wie z. B. Reliabilität und Validität der Messinstrumente genügen, da bei der Erhebung und Interpretation studentischer Evaluationsdaten zahlreiche potentielle Störfaktoren berücksichtigt werden müssen und mitunter auch weitreichende Konsequenzen aus Evaluationsergebnissen gezogen werden.

In diesem Positionspapier des GMA-Ausschusses Lehrevaluation werden strategische und methodische Aspekte der Evaluation im Kontext des Medizinstudiums und verwandter Studiengänge erörtert und – über bereits formulierte allgemeine Standards für Evaluation [4] hinaus – spezifische Empfehlungen für die Weiterentwicklung der Evaluation formuliert. Das Positionspapier soll den Fakultäten als Ausgangspunkt für die Weiterentwicklung ihrer eigenen Evaluationsprozesse dienen; gleicher-

maßen soll es dazu beitragen, die Erstellung eines standortübergreifenden Evaluationskonzeptes für die Lehre in medizinischen und verwandten Studiengängen voranzutreiben.

Hintergrund

In den Diskussionen des GMA-Ausschusses Lehrevaluation entstand der Wunsch, Fragebögen zu entwickeln, die eine umfassende studiengangs- und standortübergreifende Evaluation der Lehre ermöglichen. Auf diese Weise soll – auch vor dem Hintergrund der Überarbeitung der Nationalen Kompetenzbasierten Lernzielkataloge Medizin/Zahnmedizin und der anstehenden Reform der Approbationsordnung – eine Vergleichbarkeit zwischen den Fakultäten geschaffen werden. Im Zusammenhang mit der Digitalisierung der Lehre seit dem Sommersemester 2020 wurden in einem entsprechenden Pilotvorhaben von Mitgliedern des Ausschusses bereits verschiedene Fragebögen entwickelt, die sich sowohl an Lehrende als auch an Studierende richteten und unterschiedliche Ebenen der Lehrorganisation betrachteten (Meso- und Mikrolevel). Diese Bögen wurden allen Fakultäten zur Verfügung gestellt.
Ein standortunabhängiges Instrument wird nicht alle Dimensionen der Lehrqualität in gleicher Weise abdecken können, so dass hier eine inhaltliche Priorisierung sinnvoll erscheint. Die Konzeption eines Instruments, das universell einsetzbar ist und keine zusätzliche Hard- und Software an allen Standorten erfordert, stellt nach wie vor eine Herausforderung dar. Zudem besteht Bedarf an neuen Erhebungsinstrumenten, die aktuellen Entwicklungen (z.B. Gewicht auf Kommunikation und Interprofessionalität) und besonderen Lehr-Settings (z.B. Vermittlung praktischer Fertigkeiten, z.B. in Skills Labs) Rechnung tragen.

Grundsätzliche Überlegungen

Der Ausschuss empfiehlt das folgende vierschrittige Vorgehen:

1. **Formulieren des/der Evaluationziels/e:** Vor der Durchführung einer Evaluation muss das Ziel derselben klar benannt und konsentiert werden. Dies umfasst auch eine Formulierung des Evaluationsziels als zu messendes Konstrukt sowie Überlegungen zum Ausschluss eventueller Störvariablen. Nach Möglichkeit sollte vorab festgelegt werden, welche Konsequenzen aus den Ergebnissen gezogen werden sollen.
2. **Auswahl des/der Datenerhebungsinstruments/e:** Der zentrale Schritt im Evaluationsprozess ist die Auswahl eines oder mehrerer geeigneter Instrumente, um das genannte Ziel zu erreichen. Hier ist zu entscheiden, ob ein bestehendes Erhebungsinstrument mit bekannten psychometrischen Charakteristika gewählt werden soll oder ob eine Neuentwicklung erforderlich ist. Für spezifische Zwecke kann es sinnvoll sein, nicht-validierte Erhebungsbögen zu verwenden; generell und aus einer wissenschaftlichen Perspektive ist aber empfehlenswert, publizierte und validierte Erhebungsinstrumente einzusetzen.
3. **Implementierung der Datenerhebung:** Hier ist grundsätzlich zu klären, zu welchem Zeitpunkt bzw. in welchem Zeitraum die Daten erhoben werden sollen und ob die Datenerhebung papierbasiert oder online erfolgen soll. Im letzteren Fall stehen verschiedene Plattformen zur Verfügung, von denen einige automatisiert Benachrichtigungen und Erinnerungen an die zu befragenden Personen versenden können.
4. **Darstellung der Ergebnisse:** Bei der Darstellung der Ergebnisse sollten geeignete Methoden der Datenaufbereitung genutzt werden, die Rückschlüsse über die wesentlichen Parameter der Evaluation erlauben. Über deskriptive Auswertungen hinausgehende Analysen sollten nur in begründeten Fällen vorgenommen werden.

Je nach Zielsetzung sollte der Evaluationsprozess von den Gremien der Fakultät unterstützt werden. Zuständigkeiten, Workflows und die Umsetzung von Konsequenzen sollten durchgehend transparent kommuniziert werden.

Methodische und praktische Aspekte

Im Folgenden wird – teilweise mit Rückbezug auf die o.g. vier Schritte – auf einzelne Aspekte der Evaluationsmethodik eingegangen.

1. Formulieren des/der Evaluationziels/e

Der erste Schritt im Evaluationsprozess ist die Festlegung des Evaluationsziels. In Abhängigkeit von diesem Ziel werden passende Erhebungsinstrumente ausgewählt. Ganz allgemein besteht das Ziel der Evaluation in den meisten Fällen darin, die Qualität der Lehre zu messen und auf dieser Grundlage Verbesserungen anzustreben. Zu diesem Zweck muss zunächst bekannt sein, wie das Konstrukt „gute Lehre“ im konkreten Fall definiert ist. So war für ca. 800 klinische Dozierende an deutschen medizinischen Fakultäten in einer Umfrage “gute Lehre” insbesondere dadurch charakterisiert, dass sie bei den Studierenden zu einem nachhaltigen Lernerfolg führte und das Interesse an den gelehrten Inhalten erhöhte [5].

Zielvariable(n)

Zwei Definitionen guter Lehre – die Stanford-Kriterien und die Dimensionen nach Gibson – sollen an dieser Stelle exemplarisch näher ausgeführt werden, weil sie weithin bekannt und gut operationalisierbar sind. Außerdem können Daten, die auf Grundlage dieser Definitionen erhoben werden, konkrete Hilfestellungen bei der Weiterentwicklung der Lehre geben. Grundsätzlich existieren darüber hinaus selbstverständlich noch weitere Klassifi-

kationen und Systematiken zur Bewertung „guter Lehre“, die alle vergleichbare Dimensionen postulieren [6], [7]. Die Stanford-Kriterien [8], die ursprünglich im Rahmen eines Fakultätsentwicklungsprogramms entwickelt wurden, sind:

1. Etablierung des Lernklimas,
2. Leitung einer Lehreinheit,
3. Zielkommunikation,
4. Förderung von Verstehen und Behalten,
5. Evaluation,
6. Feedback und
7. Förderung selbstbestimmten Lernens.

Nach Gibson [2] ist Lehrqualität ein mehrdimensionales Konstrukt. Vor der Evaluation ist zu klären, welche Dimension/en guter Lehre abgebildet werden soll/en:

- Die strukturelle Dimension umfasst die Rahmenbedingungen, unter denen Lehre stattfindet, u.a. die Lehrräume, den grundsätzlichen Aufbau eines Studiengangs oder das Vorhandensein eines Lernzielkatalogs.
- Die prozedurale Dimension bezieht sich auf den Ablauf von Lehrveranstaltungen. Zu den hier gebräuchlichen Qualitätsindikatoren gehören z.B. die Pünktlichkeit der Lehrenden und Studierenden, die Art der Interaktion und der Einsatz verschiedener Lehrmethoden [9], [10]. Ein in diesem Zusammenhang häufig benutzter Begriff ist das „Lernklima“ [11].
- Die persönliche Dimension betrifft die Lehrpersonen, die den Unterricht durchführen. Lehrpersonen haben in den meisten Fällen kaum eine Möglichkeit, die strukturellen Bedingungen der Lehre zu beeinflussen. Prozedurale Aspekte der Lehre können von ihnen aber durchaus gestaltet werden. Aufgrund der hervorgehobenen Rolle der Lehrenden in den meisten Curricula wurden spezifische Erhebungsinstrumente zur Bewertung der Performanz individueller Lehrender entwickelt [12], [13], [14], [15]. Diese können auch im Rahmen von Habilitationsverfahren zum Einsatz kommen.
- Die Ergebnis-Dimension der Lehrqualität bezieht sich auf das messbare Resultat der Lehre. Dies wird häufig mit dem Lernerfolg der Studierenden gleichgesetzt (wenngleich auch andere Definitionen denkbar sind). Zur Messung des Lernerfolgs können Prüfungsleistungen herangezogen werden; diese spiegeln jedoch lediglich den Leistungsstand am Ende einer Veranstaltung(-sreihe) wider und können nichts über den tatsächlichen Zuwachs an Wissen, Fertigkeiten und Professionalität durch die Veranstaltung aussagen. Als Alternative zur Bestimmung des Lernerfolgs können auch wiederholte Selbsteinschätzungen verwendet werden, deren Reliabilität und Validität in einigen Studien untersucht wurden [16], [17], [18].

Häufig werden Studierende – ohne Bezug zu einzelnen der vorgenannten Dimensionen – um eine allgemeine Bewertung der Lehre gebeten, z.B. nach dem Schulnotenprinzip [19].
Ein solches globales Rating entspricht einem Stimmungsbild. Aufgrund der Einheitlichkeit von Schulnoten erscheinen die Ergebnisse zwischen Lehrveranstaltungen vergleichbar. Bloße Zufriedenheits-Bewertungen anhand von Schulnoten sind jedoch besonders schwer zu interpretieren, weil das Konstrukt „Zufriedenheit“ sehr individuell ist. Außerdem sind nicht alle Lehrformate, mit denen Studierende zufrieden sind, auch effektiv. Umgekehrt führen nicht alle nachweislich effektiven Lehrformate auch zu einer erhöhten Zufriedenheit bei den Studierenden.

Störvariablen

Die Validität der Messung von Zielvariablen ist bedroht, wenn konstrukt-irrelevante Faktoren die Ergebnisse beeinflussen. Das Risiko einer Verzerrung ist besonders groß, wenn das zu messende Konstrukt nicht genau genug bestimmt wurde und entsprechend auch nicht sicher zwischen (validen) Einflussgrößen und Störgrößen unterschieden werden kann.
In einer Literaturübersicht aus dem Jahr 2015 wurden zahlreiche Faktoren identifiziert, die sich auf studentische Globalbewertungen (z.B. Schulnoten-Ratings) von Lehrveranstaltungen auswirken können [20]. Während einige dieser Faktoren (z.B. Kurs-Organisation, Kommunikation und Feedback für Studierende) mit dem Konstrukt „gute Lehre“ leicht vereinbar sind, sind andere eindeutig als Störvariablen zu bezeichnen (z.B. Geschlecht der Studierenden, Interesse der Studierenden am gelehrten Inhalt, Zeitpunkt der Datensammlung). An dieser Stelle sollen zwei Aspekte besonders hervorgehoben werden:

- Format der Evaluation: Online-Datensammlungen gingen in einer Studie mit einem geringeren Rücklauf einher als papierbasierte Erhebungen [21], [22]. Zugleich fielen die Bewertungen bei der Online-Datensammlung positiver aus als bei der papierbasierten.
- Freiwilligkeit der Teilnahme an der Evaluation: Wenn die Teilnahme an der Evaluation freiwillig ist, nehmen möglicherweise eher leistungsstarke Studierende daran teil; dies kann dazu führen, dass die Bewertungen positiver ausfallen, als wenn alle Studierenden teilnehmen. Umgekehrt kann eine Verpflichtung zur Evaluation (sofern diese aus rechtlicher Sicht überhaupt möglich ist) dazu führen, dass einige Studierende die Bögen nicht ernsthaft ausfüllen und sich somit die konstrukt-irrelevante Varianz in den Daten erhöht.

Wenngleich es kaum möglich erscheint, bei der Datenanalyse stets alle Störvariablen zu berücksichtigen bzw. dafür zu adjustieren (weil aufgrund des Datenschutzes keine personenbezogenen Informationen wie z.B. das Geschlecht der Teilnehmenden erhoben werden), sollten die Personen, die Evaluationen durchführen und ihre Ergebnisse interpretieren, für diese Effekte sensibilisiert sein und sie ggf. bei der Planung der Evaluation einbeziehen.

2. Auswahl des/der Datenerhebungsinstruments/e

Der zweite Schritt im Evaluationsprozess ist die Auswahl des/r geeigneten Evaluationswerkzeugs/e. Üblicherweise werden Studierende dazu eingeladen, die Lehre zu bewerten. Alternativ oder ergänzend können aber auch Absolvent*innen oder Patient*innen darum gebeten werden, spezifische Aspekte der Lehre zu bewerten. Lehrende können mittels Selbsteinschätzungen (unter Supervision, z.B. in Fakultätsentwicklungsprogrammen) oder mittels peer ratings die Qualität ihrer eigenen Lehre bewerten. Schließlich können auch andere vorhandene Daten wie z.B. Studienverlaufsparameter oder aggregierte Prüfungsergebnisse zur Bewertung der Lehre herangezogen werden.

Entsprechend der Datenquelle und der Evaluationsziele kommen jeweils unterschiedliche Datenerhebungsformate und -instrumente zum Einsatz:

Das laut einer Umfrage von 2021 von 97% der Fakultäten eingesetzte [23] und somit gebräuchlichste Format ist der Fragebogen, der entweder papierbasiert oder online ausgefüllt werden kann. Eine Umfrage unter den Medizinischen Fakultäten in Deutschland ergab, dass an nahezu allen Standorten Online-Fragebögen eingesetzt werden. Die Hälfte der Fakultäten nutzt zusätzlich dazu papierbasierte Fragebögen [23], [24]. Für die Evaluation hochschulischer Lehre stehen zwar zahlreiche allgemeine Fragebögen zur Verfügung; medizinspezifische Aspekte wie beispielsweise der Unterricht am Krankenbett werden hier aber nicht immer angemessen berücksichtigt. Aus diesem Grund wurden spezifische Instrumente zur Bewertung in medizinischen und verwandten Studiengängen entwickelt und wissenschaftlich validiert.

Viele der an den einzelnen Fakultäten gebräuchlichen Fragebögen dienen dem Zweck, Aussagen über die Qualität der lokalen Lehre treffen zu können. Aufgrund der Spezifika einzelner Standorte, die in den Evaluationsbögen ihren Niederschlag gefunden haben, sind solche Bögen meist nicht fakultäts- oder standortübergreifend einsetzbar. Demgegenüber bieten publizierte Bögen mit bekannten psychometrischen Charakteristika den Vorteil, weitgehend standortunabhängig zu sein; sie werden jedoch bisher laut eigenen Angaben der Fakultäten nur an 7% der Standorte verwendet [23]. Kombinationen aus validierten und nicht-validierten Bögen sind im Prinzip möglich und werden auch von 30% der Fakultäten eingesetzt [23], sollten aber kenntlich gemacht werden.

Wenngleich Fragebögen Freitextfelder enthalten können, dienen die Fragebögen in den meisten Fällen dazu, quantitative Daten zu erheben. Im Unterschied dazu eignen sich Einzelinterviews oder Fokusgruppen eher dazu, qualitative Daten zu generieren. Fokusgruppen werden nach Angaben der Fakultäten in 21%, Interviews in 15% der medizinischen Studiengänge in Deutschland genutzt [24]. Hinzu kommen Kursnachbesprechungen, die mehr oder weniger strukturiert ablaufen können und in erster Linie dazu dienen, konkrete Probleme und Lösungen zu erörtern.

Bei der Konstruktion von Fragebögen mit skalierten Items sind unter anderem die Erkenntnisse aus der Forschung zur Fragebogenkonstruktion zu berücksichtigen [25], [26].

3. Implementierung der Datenerhebung

Fragen der Implementierung der Lehrevaluation beziehen sich auf die Art der Datensammlung (online, papierbasiert, persönliche oder telefonische Gespräche), auf die Verankerung im Curriculum (verpflichtende vs. freiwillige Teilnahme) und insbesondere auf den Zeitpunkt der Datensammlung. Auf Erhebungsinstrumente und eine eventuelle Verpflichtung zur Teilnahme an der Evaluation wurde bereits eingegangen. In diesem Abschnitt wird die zeitliche Verankerung der Evaluation thematisiert. Häufig findet die Evaluation am Ende einer Lehrveranstaltung, eines Kurses/Moduls oder eines Semesters/Studienabschnitts statt. Während die Lehrenden auf diese Weise den besten Überblick über die Qualität der Lehre erhalten können, kann dieser Zeitpunkt sich auf die Motivation der Studierenden negativ auswirken, da diese die Konsequenzen aus der Evaluation selbst nicht mehr erleben. Des Weiteren kann sich eine Datensammlung am Ende des Semesters durch den in vielen Fällen zu großen zeitlichen Abstand zwischen Erhebung und Gegenstand der Evaluierung negativ auf die Kriteriumsvalidität der Messung auswirken. Daher werden an einigen Standorten bereits zur Kursmitte Evaluationen abgehalten, um noch im laufenden Semester nachsteuern zu können. Zu viele Datenerhebungen in kurzer Folge können bei den Studierenden jedoch zu einer “Evaluationsmüdigkeit” führen, was sich negativ auf den Rücklauf auswirken kann. Daher wurden auch Methoden entwickelt, die es ermöglichen, bereits aus Rückmeldungen eines kleinen Teils der Studierenden-Kohorte verallgemeinerbare Rückschlüsse zu ziehen. Diese Methode wurde zunächst nur für Zufriedenheits-Bewertungen untersucht [27], [28]. Eine aktuelle Arbeit belegt aber, dass eine reliable Abschätzung des studentischen Lernerfolgs auch dann möglich ist, wenn nur ca. 20-30% der Studierenden einer Kohorte an der Evaluation teilnehmen [29].

Eine Besonderheit des Medizinstudiums ist, dass – insbesondere in klinisch geprägten Veranstaltungen – häufig nicht eine einzige Lehrperson alle Veranstaltungen eines Kurses leitet, sondern zahlreiche Personen in die Lehre kleinerer Studierendengruppen involviert sind. Diesem Umstand muss bei der Evaluation der individuellen Lehrleistung Rechnung getragen werden, z.B. indem den Studierenden ermöglicht wird, die für sie zuständigen Lehrenden zeitnah und individuell zu bewerten.

Um Störungen der Bewertungen der Lehre durch Unzufriedenheit mit der eigenen Prüfungsleistung auszuschließen, sollte eine Evaluation der Lehre vor Ablegen der Prüfung erfolgen – wenngleich das in der Realität oft nur schwer umzusetzen ist. Falls auch die Prüfung als solche einer Evaluation durch Studierende unterzogen werden

soll, ist hierfür eine gesonderte Datenerhebung einzuplanen.
Der Workflow zur Durchführung einer Evaluation innerhalb einer Fakultät gehört zwar ebenso wie die Konsequenzen aus ihren Ergebnissen im weitesten Sinne ebenfalls zur Implementierung; aufgrund der großen Bedeutung dieser Aspekte sind ihnen aber eigene Kapitel gewidmet.

4. Auswertung und Darstellung der Ergebnisse

Bei der Wahl eines geeigneten Formates für die Darstellung der Evaluationsergebnisse sind ebenfalls verschiedene Aspekte zu beachten:
Der/die Adressat*innen sollten in der Lage sein, anhand der Darstellung die Stichprobengröße, den Rücklauf, die Art der Verteilung und eventuelle Boden- und Deckeneffekte zu erkennen.
Inferenzstatistische Analysen sollten nur vorgenommen werden, wenn hierfür eine begründete Fragestellung vorliegt. In diesem Fall sollte das Verfahren benannt werden.
Außerhalb wissenschaftlicher Studien sollte angemerkt werden, dass es sich um explorative Analysen handelt.

Workflow: Datenschutz und Zuständigkeiten

Die Sicherstellung der Lehrqualität ist eine zentrale Aufgabe des Studiendekanats. Daher sollte das Studiendekanat die personelle und technische Expertise vorhalten, um Evaluationen professionell durchzuführen und die Ergebnisse innerhalb der Fakultät kommunizieren zu können. Dies umfasst zum einen technische Aspekte der Datensammlung und -aufbereitung, zum anderen aber vor allem die Kompetenz, Evaluationsziele zu beschreiben und vor dem Hintergrund der Kenntnis der aktuellen Literatur kongruente Instrumente auszuwählen.
Dezentrale Aktivitäten können sinnvoll sein (insbesondere zur Evaluation von Lehrinnovationen in umschriebenen Bereichen). Hier ist aber in besonderem Maße der Datenschutz zu beachten, wenn beispielsweise die Leistung einzelner Lehrender bewertet werden soll. Insofern – wie z.B. in diesem Fall sehr wahrscheinlich – eine Notwendigkeit zur Befassung der jeweiligen Datenschutzkommission besteht, müssen die entsprechenden Fristverläufe beachtet werden. Die Verantwortung für den Umgang mit Evaluationsdaten liegt entweder bei dem/der (Studien-)Dekan*in oder bei einer Person, die in der jeweiligen Evaluationsordnung benannt ist. Umgekehrt muss auch sichergestellt werden, dass die Personen, die sich an einer Evaluation beteiligen, anonym bleiben. Für die Erhebung von persönlichen Daten in einem Evaluationsvorhaben ist unter Umständen eine Einwilligung der Beteiligten nötig.

Konsequenzen aus Evaluationsergebnissen

Ein übergeordnetes Ziel der Evaluation der Lehre ist die Identifikation von Stärken und Schwächen und – basierend auf dieser Analyse – die kontinuierliche Verbesserung der Lehre. Die Konsequenzen aus den Evaluationsergebnissen müssen sich an den zuvor festgelegten Zielen orientieren. Sie können sich entsprechend auf die lehrbezogenen Strukturen und Prozesse, die lehrenden Personen und/oder auf das Lehrergebnis beziehen:

- Konsequenzen für die strukturelle Dimension der Lehre: Wenn sich aus der Evaluation Hinweise auf suboptimale Lehrbedingungen (Räume, Zugänglichkeit digitaler Angebote etc.) ergeben, müssen hier Verbesserungen angestrebt werden. Zu den strukturellen Bedingungen der Lehre gehört indirekt auch die finanzielle Ausstattung. Die Ableitung von Konsequenzen auf die leistungsorientierte Mittelvergabe wird jedoch kritisch diskutiert [18] und setzt zwingend voraus, dass ein qualitätsgesichertes Evaluationsinstrument eingesetzt wird und eine Vergleichbarkeit zwischen unterschiedlichen Kursen besteht.
- Konsequenzen für die prozedurale Dimension der Lehre: Je nachdem, welche Optimierungsbedarfe sich aus den Daten ergeben, könnten Maßnahmen ergriffen werden, die das Ziel verfolgen, das Lernklima zu verbessern, die Umsetzung didaktischer Formate zu optimieren oder Lernprozesse anderweitig (z.B. digital) zu unterstützen.
- Konsequenzen für persönliche Dimension der Lehre (Lehrpersonen): An vielen Standorten (82%) werden auf der Grundlage von Evaluationsdaten Feedbackgespräche mit den Lehrenden geführt [24]. Im Kontext einer personenbezogenen Evaluation können sich aus den Daten direkte individuelle Konsequenzen wie das Angebot einer weitergehenden didaktischen Schulung ergeben. Wenn der Lehre im Rahmen der Erlangung der Venia legendi ein adäquates Gewicht gegeben werden soll, kann der Verlauf des Habilitationsverfahrens von den Ergebnissen mehrzeitiger Lehrevaluationen durch verschiedene Akteur*innen (z.B. Studierende, Expert*innen für Medizindidaktik, Kolleg*innen im Sinne von critical friends) abhängig gemacht werden.
- Konsequenzen für die Ergebnis-Dimension der Lehre (messbares Resultat): Wenn die Evaluation zeigt, dass die angestrebten Lernziele nicht erreicht wurden, sollte dies umfassende Konsequenzen für das „constructive alignment“ innerhalb des betreffenden Kurses haben (d.h. sind die Lernziele angemessen/sind die Lehrmethoden adäquat/bilden die Prüfungen die Inhalte in einem dazu passenden Format ab etc.).

Empfehlungen

- Die Bewertung der Lehrqualität sollte sich auf belastbare Daten stützen. Der Ausschuss empfiehlt daher, bei der Konzeption und Implementierung lehrbezogener Evaluationen besonderen Wert auf die Kongruenz zwischen angestrebten Zielen und eingesetzten Instrumenten zur Datenerhebung zu achten.
- Um standort- und fächerübergreifende Vergleiche zu ermöglichen, sollten die Evaluationsinstrumente nach Möglichkeit validiert sein. Publizierte Evaluationsinstrumente, die Spezifika des Medizinstudiums und verwandter Studiengänge (z.B. patient*innenzentrierte Lehrformate) berücksichtigen, sind verfügbar und sollten entsprechend zum Einsatz kommen [13], [15].

Anmerkung

Das Positionspapier wurde dem GMA-Vorstand vorgelegt von diesem am 24.01.2024 verabschiedet.

ORCIDs der Autor*innen

- Nicolas Haverkamp: [0000-0001-7814-8425]
- Janina Barth: [0000-0003-1905-1257]
- Uta Dahmen: [0000-0003-3483-3388]
- Tobias Raupach: [0000-0003-2555-8097]

Interessenkonflikt

Die Autor*innen erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

Literatur

1. Kern DE, Thomas PA, Howard DM, Bass EB. Curriculum Development For Medical Education - A Six Step Approach. Baltimore/London: The Johns Hopkins University Press; 1998.
2. Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP. Enhancing Evaluation in an Undergraduate Medical Education Program. Acad Med. 2008;83(8):787-793. DOI: 10.1097/ACM.1090b
3. Berk RA. Top five flashpoints in the assessment of teaching effectiveness. Med Teach. 2013;35(1):15-26. DOI: 10.3109/0142159X.2012.732247
4. Gesellschaft für Evaluation (DeGEval). Standards für Evaluation. Mainz: DeGEval – Gesellschaft für Evaluation e.V.; 2016. Zugänglich unter/available from: https://www.degeval.org/degeval-standards/standards-fuer-evaluation/
5. Schiekirka-Schwake S, Anders S, von Steinbuchel N, Becker JC, Raupach T. Facilitators of high-quality teaching in medical school: findings from a nation-wide survey among clinical teachers. BMC Med Educ. 2017;17(1):178. DOI: 10.1186/s12909-017-1000-6
6. Kirkpatrick DL. Evaluating training programs: The four levels. San Francisco: Berrett-Koehler; 1994.
7. Rindermann H. Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts. Landau: Empirische Pädagogik e.V.; 2001.
8. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med. 1998;73(6):688-695. DOI: 10.1097/00001888-199806000-00016
9. De la Fuente J, Sander P, Fernando J, Pichardo-Martinez MC. Validation Study of the Scale for Assessment of the Teaching-Learning Process, Student Version ATLP-S. Electron J Res Educ Psychol. 2010;8(2):815-840. DOI: 10.25115/ejrep.v8i21.1397
10. Staufenbiel T. Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. Diagnostica. 2000;46(4):169-181. DOI: 10.1026//0012-1924.46.4.169
11. Roff S, McAleer S, Harden RM, Al-Qahtani M, Ahmed AU, Deza H, Groenen G, Primparyon P. Development and validation of the Dundee Ready Education Environment Measure (DREEM). Med Teach. 1997;19(4):295-299. DOI: 10.3109/01421599709034208
12. Iblher P, Zupanic M, Hartel C, Heinze H, Schmucker P, Fischer MR. The Questionnaire "SFDP26-German": a reliable tool for evaluation of clinical teaching? GMS Z Med Ausbild. 2011;28(2):Doc30. DOI: 10.3205/zma000742
13. Dreiling K, Montano D, Poinstingl H, Müller T, Schiekirka-Schwake S, Anders S, von Steinbüchel N, Raupach T. Evaluation in undergraduate medical education: Conceptualizing and validating a novel questionnaire for assessing the quality of bedside teaching. Med Teach. 2017;39(8):820-827. DOI: 10.1080/0142159X.2017.1324136
14. Zuberi RW, Bordage G, Norman GR. Validation of the SETOC instrument – Student evaluation of teaching in outpatient clinics. Adv Health Sci Educ Theory Pract. 2007;12(1):55-69. DOI: 10.1007/s10459-005-2328-y
15. Müller T, Montano D, Poinstingl H, Dreiling K, Schiekirka-Schwake S, Anders S, Raupach T, von Steinbüchel N. Evaluation of large-group lectures in medicine - development of the SETMED-L (Student Evaluation of Teaching in MEDical Lectures) questionnaire. BMC Med Educ. 2017;17(1):137. DOI: 10.1186/s12909-017-0970-8
16. Schiekirka S, Reinhardt D, Beissbarth T, Anders S, Pukrop T, Raupach T. Estimating learning outcomes from pre- and posttest student self-assessments: a longitudinal study. Acad Med. 2013;88(3):369-375. DOI: 10.1097/ACM.0b013e318280a6f6
17. Raupach T, Münscher C, Beissbarth T, Burckhardt G, Pukrop T. Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness. Med Teach. 2011;33(8):e446-453. DOI: 10.3109/0142159X.2011.586751
18. Schiekirka S, Anders S, Raupach T. Assessment of two different types of bias affecting the results of outcome-based evaluation in undergraduate medical education. BMC Med Educ. 2014;14:149. DOI: 10.1186/1472-6920-14-149
19. Schiekirka S, Feufel MA, Herrmann-Lingen C, Raupach T. Evaluation in medical education: A topical review of target parameters, data collection tools and confounding factors. Ger Med Sci. 2015;13:Doc15. DOI: 10.3205/000219
20. Schiekirka S, Raupach T. A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. BMC Med Educ. 2015;15:30. DOI: 10.1186/s12909-015-0311-8
21. Paolo AM, Bonaminio GA, Gibson C, Patridge T, Kallail K. Response rate comparisons of e-mail- and mail-distributed student evaluations. Teach Learn Med. 2000;12(2):81-84. DOI: 10.1207/S15328015TLM1202_4

22. Adams MJ, Umbach PD. Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. Res High Educ. 2012;53(5):576-591. DOI: 10.1007/s11162-011-9240-5

23. Giesler M, Kunz K. „Closing the Gap“: Ergebnisse einer Bestandsaufnahme zu Qualitätssicherungsmaßnahmen der Lehre an medizinischen Fakultäten ["Closing the Gap": Results of a survey assessing quality assurance in medical education]. Z Evid Fortbild Qual Gesundhwes. 2021;164:51-60. DOI: 10.1016/j.zefq.2021.05.008

24. Schiekirka-Schwake S, Barth J, Pfeilschifter J, Hickel R, Raupach T, Herrmann-Lingen C. National survey of evaluation practices and performance-guided resource allocation at German medical schools. Ger Med Sci. 2019;17:Doc04. DOI: 10.3205/000270

25. Albanese M, Prucha C, Barnet JH, Gjerde CL. The effect of right or left placement of the positive response on Likert-type scales used by medical students for rating instruction. Acad Med. 1997;72(7):627-630. DOI: 10.1097/00001888-199707000-00015

26. Albanese M, Prucha C, Barnet JH. Labeling each response option and the direction of the positive options impacts student course ratings. Acad Med. 1997;72(10 Suppl 1):S4-6. DOI: 10.1097/00001888-199710001-00002

27. Cohen-Schotanus J, Schonrock-Adema J, Schmidt HG. Quality of courses evaluated by 'predictions' rather than opinions: Fewer respondents needed for similar results. Med Teach. 2010;32(10):851-856. DOI: 10.3109/01421591003697465

28. Schönrock-Adema J, Lubarsky S, Chalk C, Steinert Y, Cohen-Schotanus J. 'What would my classmates say?' An international study of the prediction-based method of course evaluation. Med Educ. 2013;47(5):453-462. DOI: 10.1111/medu.12126

29. Grebener BL, Barth J, Anders S, Beissbarth T, Raupach T. A prediction-based method to estimate student learning outcome: Impact of response rate and gender differences on evaluation results. Med Teach. 2021;43(5):524-530. DOI: 10.1080/0142159X.2020.1867714

Korrespondenzadresse:
Prof. Dr. Tobias Raupach, MME
Universität Bonn, Medizinische Fakultät, Institut für Medizindidaktik, Venusberg - Campus 1, 53127 Bonn, Deutschland
tobias.raupach@ukbonn.de

Bitte zitieren als
Haverkamp N, Barth J, Schmidt D, Dahmen U, Keis O, Raupach T. Position statement of the GMA committee “teaching evaluation”. GMS J Med Educ. 2024;41(2):Doc19.
DOI: 10.3205/zma001674, URN: urn:nbn:de:0183-zma0016745

Artikel online frei zugänglich unter
https://doi.org/10.3205/zma001674

Eingereicht: *16.03.2023*
Überarbeitet: *29.11.2023*
Angenommen: *01.02.2024*
Veröffentlicht: *15.04.2024*