Evaluation in medical education: A topical review of target parameters, data collection tools and confounding factors

Abstract

Background and objective: Evaluation is an integral part of education in German medical schools. According to the quality standards set by the German Society for Evaluation, evaluation tools must provide an accurate and fair appraisal of teaching quality. Thus, data collection tools must be highly reliable and valid. This review summarises the current literature on evaluation of medical education with regard to the possible dimensions of teaching quality, the psychometric properties of survey instruments and potential confounding factors.

Methods: We searched Pubmed, PsycINFO and PSYNDEX for literature on evaluation in medical education and included studies published up until June 30, 2011 as well as articles identified in the "grey literature". Results are presented as a narrative review.

Results: We identified four dimensions of teaching quality: structure, process, teacher characteristics, and outcome. Student ratings are predominantly used to address the first three dimensions, and a number of reliable tools are available for this purpose. However, potential confounders of student ratings pose a threat to the validity of these instruments. Outcome is usually operationalised in terms of student performance on examinations, but methodological problems may limit the usability of these data for evaluation purposes. In addition, not all examinations at German medical schools meet current quality standards.

Conclusion: The choice of tools for evaluating medical education should be guided by the dimension that is targeted by the evaluation. Likewise, evaluation results can only be interpreted within the context of the construct addressed by the data collection tool that was used as well as its specific confounding factors.

Keywords: evaluation, medical education, dimension, confounder, questionnaire

Sarah Schiekirka[1]
Markus A. Feufel[2,3]
Christoph Herrmann-Lingen[4,5]
Tobias Raupach[6,7]

1 Universitätsmedizin Göttingen, Studiendekanat, Göttingen, Germany

2 Charité – Universitätsmedizin Berlin, Prodekanat für Studium und Lehre, Berlin, Germany

3 Max-Planck-Institut für Bildungsforschung, Forschungsbereich Adaptives Verhalten und Kognition und Harding Zentrum für Risikokommunikation, Berlin, Germany

4 Universitätsmedizin Göttingen, Klinik für Psychosomatische Medizin und Psychotherapie, Göttingen, Germany

5 Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, Düsseldorf, Germany

6 Universitätsmedizin Göttingen, Klinik für Kardiologie und Pneumologie, Göttingen, Germany

7 University College London, Health Behaviour Research Centre, London, Great Britain

Introduction

Medical education must meet high standards because medical school graduates – mainly physician practitioners – carry great responsibility. In order to assess the quality of education, evaluations are performed at all German medical schools. No less than 10 years ago, the German Society for Evaluation established standards for the evaluation of university level education. According to these standards, evaluation instruments must permit a fair, accurate, and reliable assessment of teaching quality [1]. Medical education differs from other study programs in that it offers restricted choice of courses and uses unique teaching formats such as problem-based learning and bedside teaching [2], [3]. Seemingly generic teaching formats (e.g., lectures) may be supplemented by elements specific to medical education (e.g., live presentations of patient case histories). Thus, it is questionable whether evaluation instruments from other study programs can readily be transferred to medical education. In general, to assess the reliability and, in particular, the validity of evaluation procedures, the construct of 'good teaching' underlying an evaluation instrument must be known. This article presents the results of a broad literature search on 'evaluation in medical education', funded by the Association of the Scientific Medical Societies in Germany (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V., AWMF). Search results were discussed by a joint committee of the AWMF and the Medizinische Fakultätentag (MFT). The literature search intended to answer the following questions:

1. Which dimensions of teaching quality can be assessed in the context of medical education?
2. Which evaluation instruments are currently used, and which outcomes do they target?
3. What are the psychometric properties of these evaluation instruments?
4. What needs to be considered when designing questionnaires for evaluation in medical education, and which confounding factors must be considered when interpreting results?

Methods

In order to address these questions, we conducted a comprehensive literature search including original research, systematic reviews, dissertations and the so-called 'grey literature' published in German or English. We searched Pubmed, PsycINFO, and PSYNDEX (keywords: 'medical education', 'undergraduate medical education', 'medical curriculum' combined with 'evaluation', 'evaluation of teaching effectiveness' and 'student ratings' and their German translations: 'Medizinische Ausbildung', 'Medizinstudium'/'Studium der Medizin', 'medizinisches Curriculum', 'Evaluation', 'Lehrevaluation', 'studentische Bewertungen') for relevant articles that have been added to the respective databases up to July 30, 2011.

Additional relevant papers were identified from reference lists of published reports. In addition, we searched the online archives of the following journals: Deutsche Medizinische Wochenschrift, GMS Zeitschrift für Medizinische Ausbildung, Hochschulmanagement, Qualität in der Wissenschaft as well as Wissenschaftsmanagement. We consulted experts in the field of medical education for recommendations of relevant articles and used Google to find additional publications. The literature was analysed until saturation was reached (i.e., until no additional content was identified with respect to the research questions).

During a second, more in-depth analysis of identified publications, we extracted those articles that provided answers to the four research questions. Content extraction was guided by a checklist prompting researchers to enter information on the dimension of teaching quality assessed as well as the data collection tool (if available along with its psychometric properties, such as Cronbach's alpha).

Results

A total of 116 articles were retrieved. Of these, 46 were found in Pubmed, 22 in PsychINFO, and 4 in PSYNDEX. In addition, 28 articles were identified in the online archives of the above-mentioned German journals. The remaining 16 articles were identified as secondary literature, by recommendation, or via Internet search engines. A complete list of all 116 articles is available in Attachment 1. Many of these articles were not specific to medical education, but focused on general issues related to evaluation of university level teaching. Furthermore, not all articles provided specific answers to the aforementioned research questions. In order to answer the first three research questions, we included 30 articles with a specific focus on medical education. With respect to the fourth research question, hardly any relevant results were identified in the literature specific to medical education. Thus, 14 additional articles without a specific focus on medical education were included. The complete list of all 116 identified articles provides information on which articles were used to answer the research questions.

Due to the broadly defined research questions and, consequently, due to the high structural and content-related heterogeneity of the identified articles, we decided to present the results in form of a narrative. This approach is currently being recommended for review articles that are mainly based on quasi-experimental studies. In this context, numerical analyses (e.g., meta-analyses) seem less well-suited to answer relevant research questions because they unnecessarily constrain the range of contents covered [4]. According to current perspectives in the field of medical didactics [5], if performed according to good scientific practice, narrative reviews may yield higher informational value than averaged figures.

The results section is organised according to the four research questions. For the first three questions, it is further structured according to the four dimensions of teaching, which are specified in the following section.

Question 1: Dimensions of teaching quality in medical education

All target parameters used to assess teaching quality described in the published literature can be categorised into four dimensions [6]: On the curricular level, structural (first) as well as procedural (second) aspects of teaching can be considered; the third quality criterion refers to teacher characteristics, and the fourth dimension refers to the outcome of teaching activities. The structural dimension comprises, for instance, the physical environment available for teaching, teaching materials as well as the design of a curriculum. The learning process refers to aspects such as teacher-student interaction or teaching/learning atmosphere. Instructor-specific characteristics include teaching skills and the level of preparation, but also the teachers' enthusiasm as perceived by their students. The outcome dimension describes aspects such as learning outcome and the development of professional attitudes as a result of teaching.

Structures and processes related to teaching are assessed by many of the published evaluation instruments (see Question 2), especially because data collection and analysis can be automated easily. A reliable and valid assessment of individual teacher performance is a more complex endeavour. The corresponding instruments must meet high psychometric standards, especially due to potential consequences of such evaluation results for individual careers.

Defining teaching quality by related outcomes appears straightforward. In this context, Blumberg [7] suggests three types of outcomes: She defines 'educational outcome' as the development of competencies for independent, life-long learning. The term 'clinical career outcomes' comprises competencies relevant for the medical profession (see also [8]). Blumberg defines 'environmental outcomes' as the development of professional attitudes towards teaching itself – in the sense that graduates consider passing on their knowledge and skills to their younger colleagues as part of their own professional role, thus shaping the environment at their teaching institutions. To date, there is no consensus on how to operationalise these three types of educational outcomes.

Question 2: Target outcomes and assessment instruments

As mentioned above, the following description of target outcomes and assessment instruments is guided by the four dimensions of teaching quality (structure, process, teacher, and outcome). Given that a clear-cut alignment between dimensions and individual instruments (and vice versa) is not always possible, we will elaborate on the available instruments in the context of the dimension primarily targeted. A summary of all identified instruments with a focus on medical education is presented in Table 1.

Educational *structures and processes* are mainly evaluated using self-administered questionnaires that are completed by students. Some of the available instruments cover both structures and processes ("Medical Student Experience Questionnaire"; MedSEQ; 32 items [9] and "Marburger Fragebogen zur Evaluation des Lehrangebots in der Medizin"; 12 items [10]). Four additional instruments focus mainly on teaching-related processes and, in this context, use the term 'learning environment'. The "Dundee Ready Education Environment Measure" (DREEM; 50 items [11]) has recently become available in German [12]. The comprehensive "Learning Environment Questionnaire" (LEQ; 65 items [13]) yields some overlap with the more concise "Measuring the School Learning Environment Survey" (MSLES; 50 items [14]).

The "Medical Instructional Quality" (MedIQ; 25 items [15]) was specifically designed for evaluating clinical teaching. It covers four aspects of clinical teaching related to outpatient settings. Among other factors, the MedIQ focuses on the clinical learning environment as well as the participation of students in patient care. A comprehensive review of additional instruments for evaluating the learning environment was published in 2010 [16].

Numerous instruments have been designed to evaluate *individual teachers* (see Table 1). Again, self-administered questionnaires are predominantly used, in most cases containing scaled items and open answer options. Most instruments specific to medical education and assessing individual teacher performance are tailored to the clinical teaching context (e.g. bedside teaching) rather than lectures and seminars. Detailed information on the available instruments can be found in Table 1. There is one noteworthy questionnaire for the assessment of teaching in outpatient settings ("Student Evaluation of Teaching in Outpatient Clinics"; SETOC [17]). Furthermore, the SFDP-26 ("Stanford Faculty Development Program" [18]) survey, which is also available in German [19] needs to be mentioned in this context. This tool was originally developed at the Mayo Clinic and, by mapping the seven "Stanford-Criteria for Good Teaching", is well-grounded in theory.

As described above, the *outcome of teaching*, i.e. student learning outcome, is reflected not only in the accumulation of knowledge and practical skills but also in the development of professional attitudes [7], [8]. Unfortunately, we did not find any instruments covering the full range of these outcomes. Some German medical schools use student performance in the written part of the second state examination as a surrogate parameter for teaching quality [20]. However, multiple-choice (MC) questions (such as those used in state examinations) mainly assess factual knowledge. By memorising the correct answer [21] or by deliberate practice of MC-questions [22], students may improve their exam results regardless of their actual knowledge. Similar limitations pertain to the Progress Test, which is used by some German medical schools. This formative assessment, which is applied re-

Table 1: Summary of all identified evaluation instruments for teaching quality

Name of assessment instrument (as cited in the literature)	# of items	# of scales	Name of scales	Scales' Cronbach's α
Dimensions: "Structures" and "Processes"				
Medical Student Experience Questionnaire (MedSEQ) [9]	32	5	1: Learning, 2: Teaching and Assessment; 3: Organization and Student Understanding of the Program, 4: Community Interaction and Value, 5: Student Support, Resources	0.63–0.80
Marburger Fragebogen zur Evaluation des Lehrangebots in der Medizin [10]	12		Lehrstruktur (structure of teaching), Lehrprozess (teaching process)	0.85–0.87
Dimension: "Processes"				
Dundee Ready Education Environment Measure (DREEM) [11]	50	5	1: Students' Perception of Teaching, 2: Students' Perception of Teachers, 3: Students' Academic Self-Perceptions, 4: Students' Perceptions of Atmosphere, 5: Students' Social Self-Perceptions	0.91
Medical Instructional Quality (MedIQ) [15]	25	4	1: The Role of the Preceptor in Facilitating Learning, 2: The Role and Context of the Clinical Environment, 3: Opportunities Available to Learn, 4: Active Involvement by the Learner in the Care of Patients	0.86–0.95
Measuring the School Learning Environment Survey (MSLES) [14]	50	7	1: Breadth of Interest, 2: Student Interaction, 3: Organization, 4: Flexibility, 5: Meaningful Learning Experience, 6: Emotional Climate, 7: Nurturance	0.70–0.85
Learning Environment Questionnaire (LEQ) [13]	65	7	1: Evaluative, 2: Academic Enthusiasm, 3: Goal Direction, 4: Authoritarianism, 5: Breadth of Interest, 6: Student Interaction, 7: Intellectual Maturity	0.74–0.87
Dimension: "Instructor"				
Maastricht Clinical Teaching Questionnaire (MCTQ) [49]	16	5	1: Modeling, 2: Coaching, 3: Exploration, 4: Articulation, 5: Safe Learning Environment	0.83–0.96
Clinical Teaching Assessment Form (CTAF) [50]	9	8	1: Knowledge and Analytical Ability, 2: Clarity and Organization, 3: Enthusiasm and Stimulation, 4: Ability to Establish Rapport, 5: Ability to Involve Student in Learning Experiences, 6: Ability to Give Direction and Feedback, 7: Ability to Demonstrate Clinical Skills and Procedures, 8: Accessibility	not reported
Student Evaluation of Teaching in Outpatient Clinics (SETOC) [17]	15	5	1: Establishing Learning Milieu, 2: Clinical Teaching, 3: General Teaching, 4: Clinical Competence, 5: Global Rating	0.90–0.98
Stanford Faculty Development Program (SFDP-26) [18]	25	7	1: Learning Climate, 2: Control of Session, 3: Communication of Goals, 4: Understanding and Retention, 5: Evaluation, 6: Feedback, 7: Self-directed Learning	0.82–0.95
SFDP-26 German [19]	25	7	1: Establishing the Learning Climate, 2: Controlling a Teaching Session, 3: Communication of Goals, 4: Encouraging Understanding and Retention, 5: Evaluation, 6: Feedback, 7: Self-directed Learning	0.69–0.92
Mayo Teaching Evaluation Form (MTEF) [51]	28	7	1: Establishing a Positive Learning Climate, 2: Control of the Teaching Session, 3: Communication of Goals, 4: Enhancing Understanding and Retention, 5: Evaluation, 6: Feedback, 7: Self-directed Learning	0.92–0.99

peatedly during the course of the curriculum, also uses MC-questions. Nevertheless, it is considered a useful and important source of information for students as well as curriculum evaluation owing to its longitudinal and cross-sectional design [23].

In general, state examinations are characterised by high internal consistency. However, learning outcomes of individual classes/courses of a given curriculum can only be assessed by analysing the exam results that were performed at the medical schools. According to a recent analysis, these exams often do not meet current quality standards [24]. Recently, an evaluation tool estimating student learning outcome from comparative self-assessments has been developed as an alternative. The tool's main advantage over end-of-course exams is its adjustment for initial student performance levels, thus facilitating a critical appraisal of the learning outcome created during a course [25].

Finally, surveys among medical school graduates can be used to assess the quality of medical education. In principle, all four dimensions of teaching quality may be measured with this method. However, the present literature search identified neither articles specific to medical education nor studies related to other types of university level teaching that systematically evaluated the quality of instruments used for this purpose.

Question 3: Psychometric properties of assessment instruments

Questionnaires as well as exam results may be analysed regarding their reliability and validity. The *reliabilities* of the instruments used to assess structural and procedural aspects of teaching are given in the last column of Table 1. Cronbach's α, signifying the lower limit of reliability, is satisfactory for most questionnaires. Interrater reliability of evaluation data depends on the numbers of completed questionnaires [26]. However, no studies have yet reported a minimum response rate that would be necessary for results to be deemed reliable (see below). Measuring the reliability of examinations is a prerequisite for using exam results for evaluation purposes. At German medical schools, however, these analyses are performed on less than 40% of summative exams [24].

A well-founded interpretation of evaluation results requires the data to be valid. While content validity of examinations and evaluation instruments is usually acceptable, data on criterion and construct validity is often lacking. In addition, confounding factors potentially impacting the validity of results need to be considered. Such factors have mainly been identified for *student ratings*, and they are being discussed below (Question 4). However, the considerations pertaining to this aspect are mainly based on literature with no direct link to medical education.

The validity of *examinations* is threatened mainly by two confounding factors [27]. Construct under-representation exists if the construct to be evaluated by the exam is not completely covered. In this case, students have an advantage if they accidentally focus their learning on those contents that are covered by the exam. The second essential confounding factor is construct-irrelevant variance. This occurs if, for instance, exam questions are constructed sub-optimally, so that the exam assesses not only obvious content knowledge but also students' abilities to cope with questions that are difficult to understand. Due to a lack of valid external criteria and necessary resources, criterion validity of examinations is usually not evaluated. The above-mentioned instrument for calculating student learning outcomes from comparative self-assessments has been shown to be construct-valid in a first study [25]. Additional published results were not available at the time of the literature search. Similarly, we did not identify any studies on the reliability and validity of graduate surveys.

Question 4: Questionnaire design and confounding factors

The most common evaluation instrument in practice as well as in the identified publications is the self-administered questionnaire. When designing and using questionnaires, several aspects must be considered. As mentioned above, hardly any articles addressing this question were identified. Thus, below we present some of the pertinent findings related to questionnaire design and the most important confounding factors of self-administered evaluation instruments, mainly without a direct link to medical education.

Question type, scale options and data collection procedures may all impact on the psychometric properties of questionnaires. With respect to question type, there are open questions and scaled items. Free-text comments can yield valuable qualitative information, but not every student volunteers their opinion. Scaled items lend themselves to quantitative analyses. Global ratings that are frequently used to obtain an overall appraisal of a course (e.g., using school grades) are criticized by some authors due to their susceptibility to confounding (see below) [28], [29]. Other authors contend that the construct of good teaching is virtually one-dimensional and thus can well be assessed using global ratings [30]. Additional studies show that the reliability of instruments is positively related to the number of specific items contained [31], [32].

Scaled questions yield more favourable ratings if the positive anchor is placed on the left [33]. Furthermore, the wording of items may be interpreted differently by individual students [3]. In addition, the evaluation procedure itself needs to be considered. This factor becomes increasingly important because many medical schools have moved their evaluations to online platforms. In general, online evaluations yield lower response rates than traditional paper-based evaluations. While one study did not demonstrate an effect of this on evaluation results (in fact, students provided even more comments on the online version) [34], another report stated that low-performing students were less likely to participate in online evaluations than their high-performing peers [35]. In ad-

dition, anonymous evaluations typically yield less favourable ratings than evaluations requiring students to provide identifying information [36]. With respect to graduate surveys, it should be considered that evaluation results tend to get worse the more time has passed between exposure to teaching and data collection [37].

Items that are used to evaluate individual teachers are particularly prone to confounding. It has been shown that teachers who are enthusiastic and who have a good reputation systematically receive more favourable ratings [38], even if the content they present is flawed [39], [40]. Another important confounding factor is student interest in a course [41], [42]: Courses with voluntary participation typically receive more positive ratings than compulsory courses [28], [43]. Moreover, well-attended courses are generally evaluated more positively [44]. In the context of medical education, teaching in subjects related to basic science and theoretical medicine tend to receive less favourable ratings than clinical teaching. Similarly, lectures yield worse evaluations than small-group formats [37].

Discussion

The present article is a broad review of the available literature on evaluation in medical education. The results suggest that teaching quality is not a univariate construct. Rather, all four – partially overlapping – dimensions ('structure', 'process', 'instructor', and 'outcome') can and should be considered in evaluations. In addition, interpretation of evaluation results needs to be informed by the construct underlying the data collection tool. For instance, student appraisals of a teacher's punctuality or the condition of classrooms do not allow direct conclusions to be drawn on student learning outcomes. Exam results may be used to estimate learning outcome. However, they merely reflect performance at one point in time and do not provide information on progress during a course. Progress testing is one solution to this problem, but given that it is solely based on multiple choice questions it is unable to assess practical skills or professional attitudes. In addition, it does not use a pre-post design, which would be necessary to evaluate individual courses or modules (as opposed to student cohorts or entire study programs).

The quantitative analysis of evaluation data (e.g., by calculating means of global course ratings provided by students using a grading system) facilitates comparisons across courses. However, this approach entails two risks: First, global ratings are unlikely to represent a clear-cut construct. Second, such ratings are prone to several confounding factors [45]. If one assumes that teaching quality with all its facets can be reflected by one single mean rating, both risks threaten the reliability and validity of such global assessments. In addition to the confounding factors mentioned above, the length of data collection tools should be mentioned at this point. Some of the questionnaires listed in Table 1 contain more than 60 items and are probably not well-suited for frequent and regular use in course evaluations due to low student acceptance [46].

Less than half of the articles identified in the initial search were included in this review. The main reason for exclusion was a lack of relatedness to medical education. For instance, the validated questionnaire SEEQ ("Students' Evaluation of Educational Quality") [47] is widely used in higher education institutions in the United States. It is unclear to which extent this instrument can be generalized to medical education as its items are not specific for medical education. In addition, this questionnaire was developed for higher education in the U.S. which differs from the German setting in some respect. German instruments used to evaluate (non-medical) teaching are the HILVE ("Heidelberger Inventar zur Lehrveranstaltungs-Evaluation") [48] and the HILVE II. Both tools possess good psychometric properties, but again generalisability to medical education is questionable. Due to the specifics of medical education mentioned above, further psychometric testing is definitely advisable before applying this tool.

The results of this literature review do not justify general recommendations to be made for the use of specific questionnaires to evaluate medical education in Germany. One reason for this is that the choice of the data collection tool should be guided by the goal of evaluation. However, a preliminary and resource efficient solution could be to use the Marburger questionnaire (for structural and procedural aspects) and the SFDP-26 German [19] (for teachers), as they are already available in German and possess good psychometric characteristics. Since those instruments that were mainly developed and validated in English-speaking countries cannot easily be transferred to the context of medical education in Germany, a medium-term goal should be to design a new questionnaire from existing and new items and validate this new tool in German medical schools. This process should be informed by psychometric expertise and could involve several German medical schools as part of a related research project. By using an instrument that has been mutually agreed upon at multiple locations, greater comparability of the results could be achieved. A possible development and implementation strategy is currently being discussed between MFT and AWMF.

There is a risk that relevant publications have not been included in our final selection of papers for this review. The main limitation of the present article is that the majority of included studies were done in English-speaking countries where medical education can differ substantially from Germany (e.g., clerkships cannot readily be compared to the German 'Blockpraktikum' and 'Famulatur'; there is no direct equivalent to the 'Praktische Jahr' in most English-speaking countries). In addition, the sources used for answering the fourth research question were largely not specific to medical education. At best, it is questionable if the insights into questionnaire design and confounding factors as they pertain to evaluation in other disciplines can readily be transferred to medical education. Finally, our search for published instruments used

to assess teaching quality mainly identified self-administered questionnaires that are completed by students. Other data collection procedures (e.g., graduate surveys) might also provide helpful information. Due to limited data, we chose not to discuss these instruments in the present review.

Conclusion

The evaluation of medical education is mainly based on student ratings of structural and procedural aspects of teaching as well as the performance of individual teachers. The present review identified several reliable instruments to assess these three dimensions of teaching quality. However, evaluation research unrelated to medicine has identified a number of confounding factors impacting on student ratings, thereby threatening the validity of these instruments. These confounding factors should be considered or re-addressed when using student ratings to evaluate medical education. In Germany, the assessment of teaching quality based on exam performance is problematic as there is currently no comprehensive quality control of summative exams at German medical schools. Graduate surveys are not widely used and rely on instruments with unknown validity and reliability.

Clinical and practical implications

- The quality of medical education is a multi-dimensional construct; the four basic dimensions for assessing teaching quality are structures, processes, teacher characteristics, and learning outcome.
- To assess structures, processes and individual teachers in medical education, several instruments with good psychometric characteristics are available. The assessment of learning outcome is limited mainly due to unknown or insufficient reliability and validity of summative exams in medical schools.
- When designing and implementing evaluation instruments, the confounding factors presented in this review must be taken into account as far as they are likely to generalise from other fields of university level teaching to medical education.

Notes

Competing interests

The authors declare that they have no competing interests.

Authorship

The authors Herrmann-Lingen C and Raupach T contributed equally to this work.

Attachments

Available from
http://www.egms.de/en/journals/gms/2015-13/000219.shtml

1. 000219_Appendix.pdf (151 KB)
   Complete list of the literature

References

1. DeGEval – Gesellschaft für Evaluation e.V., editor. Standards für Evaluation. Köln: DeGEval; 2002.
2. Kogan JR, Shea JA. Course evaluation in medical education. Teach Teach Educ. 2007;23(3):251-64. DOI: 10.1016/j.tate.2006.12.020
3. Billings-Gagliardi S, Barrett SV, Mazor KM. Interpreting course evaluation results: insights from thinkaloud interviews with medical students. Med Educ. 2004 Oct;38(10):1061-70. DOI: 10.1111/j.1365-2929.2004.01953.x
4. Colliver JA, Kucera K, Verhulst SJ. Meta-analysis of quasi-experimental research: are systematic narrative reviews indicated? Med Educ. 2008 Sep;42(9):858-65. DOI: 10.1111/j.1365-2923.2008.03144.x
5. Eva KW. On the limits of systematicity. Med Educ. 2008 Sep;42(9):852-3. DOI: 10.1111/j.1365-2923.2008.03140.x
6. Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP. Enhancing evaluation in an undergraduate medical education program. Acad Med. 2008 Aug;83(8):787-93. DOI: 10.1097/ACM.0b013e31817eb8ab
7. Blumberg P. Multidimensional outcome considerations in assessing the efficacy of medical educational programs. Teach Learn Med. 2003;15(3):210-4. DOI: 10.1207/S15328015TLM1503_10
8. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach. 2007 Sep;29(7):642-7. DOI: 10.1080/01421590701746983
9. Boyle P, Grimm MC, McNeil HP, Scicluna H. The UNSW Medicine Student Experience Questionnaire (MedSEQ). San Francisco: Academia; 2009. Available from: http://www.academia.edu/5252480/Medicine_Student_Experience_Questionnaire_MEDSEQ_UNSW
10. Krebs K. Marburger Fragebogen zur Evaluation des Lehrangebots in der Medizin: Eine Untersuchung zur Reliabilität und Dimensionalität des Marburger Fragebogens zur Evaluation des Lehrangebots am Fachbereich Medizin [Dissertation]. Marburg: Philipps-Universität Marburg; 2006. Available from: http://archiv.ub.uni-marburg.de/diss/z2006/0387/pdf/dkk.pdf
11. Roff S. The Dundee Ready Educational Environment Measure (DREEM)– a generic instrument for measuring students' perceptions of undergraduate health professions curricula. Med Teach. 2005 Jun;27(4):322-5. DOI: 10.1080/01421590500151054
12. Rotthoff T, Ostapczuk MS, De Bruin J, Decking U, Schneider M, Ritz-Timme S. Assessing the learning environment of a faculty: psychometric validation of the German version of the Dundee Ready Education Environment Measure with students and teachers. Med Teach. 2011;33(11):e624-36. DOI: 10.3109/0142159X.2011.610841
13. Rothman AI, Ayoade F. The development of a learning environment: a questionnaire for use in curriculum evaluation. J Med Educ. 1970;45(10):754-9. DOI: 10.1097/00001888-197010000-00006

14. Marshall RE. Measuring the medical school learning environment. Acad Med. 1978;53(2):98-104. DOI: 10.1097/00001888-197802000-00003

15. James PA, Osborne JW. A measure of medical instructional quality in ambulatory settings: the MedIQ. Fam Med. 1999 Apr;31(4):263-9.

16. Soemantri D, Herrera C, Riquelme A. Measuring the educational environment in health professions studies: a systematic review. Med Teach. 2010;32(12):947-52. DOI: 10.3109/01421591003686229

17. 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 Feb;12(1):55-69. DOI: 10.1007/s10459-005-2328-y

18. 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):688-95. DOI: 10.1097/00001888-199806000-00016

19. Iblher P, Zupanic M, Härtel C, Heinze H, Schmucker P, Fischer MR. Der Fragebogen "SFDP26-German": Ein verlässliches Instrument zur Evaluation des klinischen Unterrichts? [The Questionnaire "SFDP26-German": a reliable tool for evaluation of clinical teaching?]. GMS Z Med Ausbild. 2011;28(2):Doc30. DOI: 10.3205/zma000742

20. Herzig S, Marschall B, Nast-Kolb D, Soboll S, Rump LC, Hilgers RD. Positionspapier der nordrhein-westfälischen Studiendekane zur hochschulvergleichenden leistungsorientierten Mittelvergabe für die Lehre [Distribution of government funds according to teaching performance]. GMS Z Med Ausbild. 2007;24(2):Doc109. Available from: http://www.egms.de/en/journals/zma/2007-24/zma000403.shtml

21. Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen [Format and implementation of written assessments]. GMS Z Med Ausbild. 2006; 23(3):Doc44. Available from: http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml

22. Mahamed A, Gregory PA, Austin Z. "Testwiseness" among international pharmacy graduates and Canadian senior pharmacy students. Am J Pharm Educ. 2006 Dec;70(6):131. DOI: 10.5688/aj7006131

23. Freeman A, Van Der Vleuten C, Nouns Z, Ricketts C. Progress testing internationally. Med Teach. 2010;32(6):451-5. DOI: 10.3109/0142159X.2010.485231

24. Möltner A, Duelli R, Resch F, Schultz JH, Jünger J. Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten [School-specific assessment in German medical schools]. GMS Z Med Ausbild. 2010;27(3):Doc44. DOI: 10.3205/zma000681

25. 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):e446-53. DOI: 10.3109/0142159X.2011.586751

26. Spiel C, Schober B, Reimann R. Evaluation of curricula in higher education: challenges for evaluators. Eval Rev. 2006 Aug;30(4):430-50. DOI: 10.1177/0193841X05285077

27. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327-33. DOI: 10.1046/j.1365-2923.2004.01777.x

28. Aleamoni LM. Student rating myths versus research facts from 1924 to 1998. J Pers Eval Educ. 1999;13(2):153-66. DOI: 10.1023/A:1008168421283

29. Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. Am Psychol. 1997;52(11):1187-97. DOI: 10.1037/0003-066X.52.11.1187

30. d'Apollonia S, Abrami PC. Navigating student ratings of instruction. Am Psychol. 1997;52(11):1198-208. DOI: 10.1037/0003-066X.52.11.1198

31. Jackson DL, Teal CR, Raines SJ, Nansel TR, Force RC, Burdsal CA. The dimensions of students' perceptions of teaching effectiveness. Educ Psychol Meas. 1999;59(4):580-96. DOI: 10.1177/00131649921970035

32. Marsh HW. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. J Educ Psychol. 1984;76(5):707-54. DOI: 10.1037/0022-0663.76.5.707

33. 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 Jul;72(7):627-30. DOI: 10.1097/00001888-199707000-00015

34. Sorenson DL, Johnson TD. Online student ratings of instruction. New Dir Teach Learn. 2003;2003(96):1-112.

35. 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):576-91. DOI: 10.1007/s11162-011-9240-5

36. Afonso NM, Cardozo LJ, Mascarenhas OA, Aranha AN, Shah C. Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. Fam Med. 2005 Jan;37(1):43-7.

37. van den Bussche H, Weidtmann K, Kohler N, Frost M, Kaduszkiewicz H. Evaluation der ärztlichen Ausbildung: Methodische Probleme der Durchführung und der Interpretation von Ergebnissen [Evaluation of medical education: methodological problems of implementation and interpretation of results]. GMS Z Med Ausbild. 2006;23(2):Doc37. Available from: http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml

38. Griffin BW. Instructor reputation and student ratings of instruction. Contemp Educ Psychol. 2001 Oct;26(4):534-52. DOI: 10.1006/ceps.2000.1075

39. Marsh HW, Ware JE. Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. J Educ Psychol. 1982;74(1):126-34. DOI: 10.1037/0022-0663.74.1.126

40. Naftulin DH, Ware JE, Donnelly FA. The Doctor Fox Lecture: a paradigm of educational seduction. J Med Educ. 1973 Jul;48(7):630-5. DOI: 10.1097/00001888-197307000-00003

41. Prave RS, Baril GL. Instructor ratings: Controlling for bias from Initial student interest. J Educ Bus. 1993;68(6):362-6. DOI: 10.1080/08832323.1993.10117644

42. Cashin WE. Student ratings of teaching: A summary of the research. East Lansing, MI, USA: Office of Faculty and Organizational Development at Michigan State University; 1988. (IDEA Paper; No.20). Available from: http://ideaedu.org/wp-content/uploads/2014/11/idea-paper_50.pdf

43. Ting KF. A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. Res High Educ. 2000;41(5):637-61. DOI: 10.1023/A:1007075516271

44. Abrami PC, D'Apollonia S, Cohen PA. Validity of student ratings of instruction: What we know and what we do not. J Educ Psychol. 1990;82(2):219-31. DOI: 10.1037/0022-0663.82.2.219

45. Schiekirka S, Raupach T. A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. BMC Med Educ. 2015 Mar 5;15:30. DOI: 10.1186/s12909-015-0311-8

46. Schiekirka S, Reinhardt D, Heim S, Fabry G, Pukrop T, Anders S, Raupach T. Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. BMC Med Educ. 2012 Jun 22;12:45. DOI: 10.1186/1472-6920-12-45

47. Marsh HW. SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. Brit J Psychol. 1982;52(1):77-95. DOI: 10.1111/j.2044-8279.1982.tb02505.x

48. Rindermann H, Schofield N. Generalizability of Multidimensional Student Ratings of University Instruction Across Courses and Teachers. Res High Educ. 2001;42(4):377-99. DOI: 10.1023/A:1011050724796

49. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. Acad Med. 2010 Nov;85(11):1732-8. DOI: 10.1097/ACM.0b013e3181f554d6

50. Irby DM, Gillmore GM, Ramsey PG. Factors affecting ratings of clinical teachers by medical students and residents. Acad Med. 1987;62(1):1-7. DOI: 10.1097/00001888-198701000-00001

51. Beckman TJ, Lee MC, Rohren CH, Pankratz VS. Evaluating an instrument for the peer review of inpatient teaching. Med Teach. 2003 Mar;25(2):131-5. DOI: 10.1080/0142159031000092508

Corresponding author:
Dipl.-Psych. Sarah Schiekirka
Universitätsmedizin Göttingen, Studiendekanat, Humboldtallee 38, 37073 Göttingen, Germany, Phone: +49-(0)551-39-12302, Fax: +49-(0)551/39-13012302
sarah.schiekirka@med.uni-goettingen.de

Please cite as
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. GMS Ger Med Sci. 2015;13:Doc15.
DOI: 10.3205/000219, URN: urn:nbn:de:0183-0002197

This article is freely available from
http://www.egms.de/en/journals/gms/2015-13/000219.shtml

Received: *2015-04-02*
Revised: *2015-08-31*
Published: *2015-09-16*

Evaluation im Medizinstudium: Zielgrößen, Erhebungsinstrumente und Störfaktoren – eine Annäherung

Sarah Schiekirka[1]
Markus A. Feufel[2,3]
Christoph Herrmann-Lingen[4,5]
Tobias Raupach[6,7]

1 Universitätsmedizin Göttingen, Studiendekanat, Göttingen, Deutschland
2 Charité – Universitätsmedizin Berlin, Prodekanat für Studium und Lehre, Berlin, Deutschland
3 Max-Planck-Institut für Bildungsforschung, Forschungsbereich Adaptives Verhalten und Kognition und Harding Zentrum für Risikokommunikation, Berlin, Deutschland
4 Universitätsmedizin Göttingen, Klinik für Psychosomatische Medizin und Psychotherapie, Göttingen, Deutschland
5 Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, Düsseldorf, Deutschland
6 Universitätsmedizin Göttingen, Klinik für Kardiologie und Pneumologie, Göttingen, Deutschland
7 University College London, Health Behaviour Research Centre, London, Großbritannien

Zusammenfassung

Hintergrund und Fragestellung: Die Evaluation ist fester Bestandteil der Lehre an Medizinischen Fakultäten. Gemäß den Standards der Deutschen Gesellschaft für Evaluation müssen Evaluationsinstrumente eine faire und genaue Beurteilung der Lehrqualität erlauben. Entsprechend müssen die genutzten Erhebungsinstrumente eine hohe Reliabilität und Validität aufweisen. In dieser Übersichtsarbeit wird die verfügbare Literatur zur Evaluation des Medizinstudiums mit Hinblick auf die möglichen Dimensionen der Lehrqualität, die psychometrischen Eigenschaften der Instrumente und potentielle Störfaktoren dargestellt.
Methoden: Ausgehend von einer Schlagwortsuche in Pubmed, PsycINFO und PSYNDEX wurde eine Literatur-Recherche zur Evaluation im Medizinstudium durchgeführt. Berücksichtigung fanden Arbeiten, die bis zum 30.6.2011 in die Datenbanken aufgenommen wurden sowie „graue Literatur“. Die Ergebnisse werden in narrativer Form präsentiert.
Ergebnisse: Es wurden vier Dimensionen der Lehrqualität im Medizinstudium identifiziert: Strukturen, Prozesse, Dozenten-Charakteristika und das Lehr-Ergebnis. Zur Betrachtung der ersten drei Dimensionen werden in erster Linie studentische Bewertungen herangezogen. Hierfür liegen einige reliable, in deutscher Sprache verfügbare Instrumente vor. Die Validität studentischer Bewertungen wird jedoch durch zahlreiche potentielle Störfaktoren eingeschränkt. Zur Beurteilung des Lehr-Ergebnisses werden vor allem Prüfungsleistungen herangezogen, deren Nutzbarkeit allerdings aufgrund methodischer Probleme eingeschränkt sein kann. Zudem genügen nicht alle Prüfungen an deutschen medizinischen Fakultäten den gängigen Qualitätsstandards.
Folgerung: Die Auswahl von Instrumenten zur Evaluation des Medizinstudiums sollte sich daran orientieren, welche Dimension der Lehre beurteilt werden soll. Entsprechend können Evaluationsergebnisse auch nur vor dem Hintergrund des vom genutzten Erhebungsinstrument abgebildeten Konstrukts und dessen spezifischen Störfaktoren interpretiert werden.

Schlüsselwörter: Evaluation, Medizinstudium, Dimension, Störfaktor, Fragebogen

Einleitung

Das Medizinstudium muss höchsten Qualitätsstandards genügen, da die Absolventen medizinischer Fakultäten – in erster Linie Ärztinnen und Ärzte – große Verantwortung tragen. Zur Bewertung der Qualität der Lehre werden an allen deutschen medizinischen Fakultäten Evaluationen durchgeführt. Die Deutsche Gesellschaft für Evaluation hat bereits vor über zehn Jahren Standards für die Evaluation der Hochschullehre festgelegt. Diesen zufolge müssen Evaluationsinstrumente eine faire, genaue und verlässliche Beurteilung der Lehrqualität erlauben [1]. Auch ist zu berücksichtigen, dass das Medizinstudium einige Besonderheiten gegenüber anderen Studiengängen aufweist [2], [3], beispielweise wenig Freiheiten bezüglich der Kurswahl sowie spezifische Unterrichtsformen wie das Problem-orientierte Lernen (POL) oder der Unterricht am Krankenbett (UaK). Selbst in scheinbar allgemeintypischen Veranstaltungstypen wie Vorlesungen können Besonderheiten wie Patientenvorstellungen auftreten. Somit ist fraglich, ob Evaluationsinstrumente aus anderen Studiengängen problemlos auf die Lehre im Medizinstudium übertragbar sind. Grundsätzlich muss zur Beurteilung der Reliabilität und insbesondere der Validität der eingesetzten Verfahren zunächst bekannt sein, welches Konstrukt von „guter Lehre“ einem Evaluationsinstrument zugrunde liegt. In der vorliegenden Arbeit werden die Ergebnisse einer breit angelegten Literaturrecherche zum Thema „Evaluation im Studium der Humanmedizin“ vorgestellt, die von der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) finanziert wurde. Die Ergebnisse wurden im Anschluss an die Recherche in einem gemeinsamen Gremium der AWMF und des Medizinischen Fakultätentags diskutiert. Im Rahmen der Literaturrecherche sollten folgende Leitfragen beantwortet werden:

1. In welchen Dimensionen kann die Qualität der medizinischen Lehre erfasst werden?
2. Welche Instrumente kommen derzeit zum Einsatz und welche Zielgrößen werden von ihnen betrachtet?
3. Welche psychometrischen Eigenschaften besitzen diese Erhebungsinstrumente?
4. Was ist bei der Konstruktion von Fragebögen für die Evaluation im Medizinstudium zu beachten, und welche Störgrößen müssen bei der Interpretation der Ergebnisse berücksichtigt werden?

Methoden

Zur Beantwortung der genannten Forschungsfragen erfolgte eine breit angelegte Literaturrecherche, in die publizierte Original- und Übersichtsarbeiten, Dissertationen sowie so genannte „Graue Literatur“ in deutscher und englischer Sprache einbezogen wurden. In einer Schlagwortsuche in Pubmed, PsycINFO und PSYNDEX (Begriffe: „medical education“, „undergraduate medical education“, „medical curriculum“ kombiniert mit „evaluation“, „evaluation of teaching effectiveness“ und „student ratings“ bzw. die analogen deutschen Begriffe: „Medizinische Ausbildung“, „Medizinstudium“/„Studium der Medizin“, „medizinisches Curriculum“, „Evaluation“, „Lehrevaluation“, „studentische Bewertungen“) wurden zunächst relevante Arbeiten identifiziert, die seit Beginn der Erfassung in den jeweiligen Datenbanken bis zum 30.6.2011 publiziert wurden.

Eine Durchsicht der Literaturverzeichnisse dieser Arbeiten lieferte Hinweise auf weitere relevante Beiträge. Des Weiteren fand eine Suche direkt in den Online-Archiven folgender Zeitschriften statt: Deutsche Medizinische Wochenschrift, GMS Zeitschrift für Medizinische Ausbildung, Hochschulmanagement, Qualität in der Wissenschaft sowie Wissenschaftsmanagement. Ebenfalls wurde persönlichen Literaturempfehlungen von Experten auf dem Gebiet der medizinischen Ausbildungsforschung gefolgt und mittels der allgemeinen Internetsuchmaschine Google gesucht. Die Literatur wurde im Hinblick auf die Forschungsfragen gesichtet, bis eine inhaltliche Sättigung erreicht war (d.h. bis keine neuen inhaltlichen Aspekte mehr identifiziert werden konnten).

Aus den identifizierten Publikationen wurden in einer zweiten, tiefergehenden Durchsicht diejenigen Arbeiten extrahiert, aus denen Antworten auf die vier oben genannten Forschungsfragen abgeleitet werden konnten. Die inhaltliche Extraktion erfolgte anhand einer Vorlage, in die jeweils die betrachtete Dimension der Lehrqualität und das genutzte Datenerhebungsinstrument (falls verfügbar mitsamt den psychometrischen Eigenschaften, insbesondere Cronbach's α) eingetragen wurde.

Ergebnisse

Insgesamt wurden 116 Arbeiten gefunden, davon 46 in Pubmed, 22 in PsycINFO und vier in PSYNDEX. Des Weiteren konnten 28 Arbeiten in Online-Archiven der oben genannten deutschen Zeitschriften, identifiziert werden. Die übrigen 16 Arbeiten wurden als Sekundärliteratur, Empfehlung oder durch allgemeine Internetsuchmaschinen gefunden. Eine komplette Liste der 116 Artikel ist in Anhang 1 verfügbar. Viele dieser Arbeiten waren jedoch nicht auf die medizinische Lehre bezogen, sondern enthielten eher allgemeine, für die Evaluation in der Hochschullehre relevante, Betrachtungen. Zudem fanden sich nicht in allen Arbeiten konkrete Antworten auf die oben formulierten Forschungsfragen. Zur Beantwortung der ersten drei Forschungsfragen wurde auf die Inhalte derjenigen 30 Volltext-Arbeiten zurückgegriffen, die einen direkten Bezug zum Medizinstudium aufwiesen. Hinsichtlich der vierten Forschungsfrage fanden sich in der medizinspezifischen Literatur kaum verwertbare Ergebnisse, so dass zu diesem Punkt auch die nicht-medizinspezifische Literatur einbezogen wurde (14 weitere Arbeiten). In der Komplettliste der 116 Volltext-Artikel wurden die Artikel gekennzeichnet, die zur Beantwortung der einzelnen Forschungsfragen herangezogen wurden.

Aufgrund der inhaltlich breit angelegten Forschungsfragen und folglich hohen inhaltlichen und strukturellen Heterogenität der eingeschlossenen Arbeiten entschieden wir uns für eine narrative Darstellung der Ergebnisse. Dieses Vorgehen wird aktuell für Übersichten empfohlen, in denen überwiegend quasi-experimentelle Studien berücksichtigt werden. In diesem Kontext erscheinen numerische Auswertungsverfahren (z.B. Meta-Analysen) zur Bearbeitung entsprechender Fragestellungen nicht optimal, da hierdurch das Spektrum der abgedeckten Inhalte unnötig eingeengt wird [4]. Nach aktueller medizindidaktischer Lehrmeinung [5] können narrative Übersichten bei guter wissenschaftlicher Durchführung einen höheren Informationsgehalt bieten als gemittelte Kennzahlen.
Die Präsentation orientiert sich an den vier oben genannten Fragen und wird innerhalb der ersten drei Forschungsfragen nach den vier verschiedenen Dimensionen der Lehre gegliedert, die im Folgenden genauer dargestellt werden.

Frage 1: Qualitäts-Dimensionen der medizinischen Hochschullehre

Alle in der publizierten Literatur beschriebenen Zielparameter zur Bewertung der Lehrqualität lassen sich einer von vier Dimensionen zuordnen [6]: Auf curricularer Ebene können sowohl strukturelle als auch prozedurale Kenngrößen der Lehre betrachtet werden; als drittes Qualitätskriterium stehen Dozenten-spezifische Charakteristika, als vierte Dimension das Ergebnis der Lehre zur Verfügung. Die Strukturdimension umfasst beispielweise die räumliche Ausstattung der Lehre, Arbeitsmaterialien sowie die Konzeption des Studiums. Der Lehrprozess meint Aspekte wie Interaktion oder Lehr-/Lernatmosphäre. Dozentenspezifische Charakteristika können unter anderem das didaktische Geschick sowie die Vorbereitung, aber auch der von den Studierenden wahrgenommene Enthusiasmus von Lehrenden sein. Die Ergebnisdimension beschreibt Aspekte wie den Lernerfolg und Entwicklung professioneller Einstellungen durch die Lehre.
Lehrbezogene Strukturen und Prozesse werden von vielen publizierten Evaluationsinstrumenten erfasst (siehe Frage 2), zumal die Datenerhebung und -auswertung leicht automatisierbar ist. Die reliable und valide Bewertung der Lehrleistung individueller Dozenten ist weitaus komplexer. Insbesondere aufgrund möglicher Konsequenzen solcher Evaluationsergebnisse für die Karriere wissenschaftlicher Mitarbeiter müssen die entsprechenden Instrumente besonders hohen psychometrischen Ansprüchen genügen.
Eine Beurteilung der Lehrqualität anhand der im Rahmen der Lehre erzielten Ergebnisse erscheint intuitiv. Blumberg [7] schlägt diesbezüglich drei Qualitäten vor: Als „educational outcome“ bezeichnet die Autorin die Entwicklung von Fertigkeiten zum eigenständigen lebenslangen Lernen. Unter „clinical career outcomes“ werden die für den Arztberuf erforderlichen Kompetenzen zusammengefasst (siehe auch [8]). Unter „environmental outcomes“ versteht Blumberg die Ausbildung einer professionellen Einstellung zur Lehre selbst – in dem Sinne, dass Absolventen die Weitergabe von Wissen und Fertigkeiten als eigene professionelle Aufgabe im Beruf verstehen und somit das Klima an Ausbildungsstätten prägen. Bislang fehlt aber ein allgemein anerkanntes Konzept zur Operationalisierung dieser “educational outcomes“.

Frage 2: Zielgrößen und Erhebungsinstrumente

Die folgende Darstellung der Zielgrößen und Erhebungsinstrumente orientiert sich wie oben bereits erwähnt an vier Dimensionen der Lehrqualität: Struktur, Prozess, Dozent und Ergebnis. Da eine trennscharfe Zuordnung der vier Dimension zu den einzelnen Instrumenten (und umgekehrt) nicht immer möglich ist, werden die verfügbaren Instrumente im Kontext derjenigen Dimension erörtert, auf die sie in erster Linie abzielen. Eine Zusammenschau aller identifizierten medizinspezifischen Instrumente bietet Tabelle 1.
Lehrbezogene *Strukturen und Prozesse* werden vorrangig mit Hilfe von Fragebögen evaluiert, die von den Studierenden selbst ausgefüllt werden. Einige der verfügbaren Instrumente decken sowohl Strukturen als auch Prozesse ab („Medical Student Experience Questionnaire“; MedSEQ; 32 Items [9] und „Marburger Fragebogen zur Evaluation des Lehrangebots in der Medizin“; 12 Items [10]). Vier weitere Instrumente beziehen sich in erster Linie auf lehrbezogene Prozesse und verwenden in diesem Kontext den Begriff „Lernumgebung“ („learning environment“). Das „Dundee Ready Education Environment Measure“ (DREEM; 50 Items [11]) steht seit kurzem auch auf Deutsch zur Verfügung [12]. Der sehr umfassende „Learning Environment Questionnaire“ (LEQ; 65 Items [13]) weist Überschneidungen mit dem etwas kürzeren „Measuring the School Learning Envionment Survey“ (MSLES; 50 Items [14]) auf.
Speziell für die Evaluation der klinischen Lehre wurde das Instrument „Medical Instructional Quality“ (MedIQ; 25 Items [15]) entwickelt, das vier Aspekte der klinischen Lehre im ambulanten Setting erfasst; unter anderem werden hier das klinische Lernumfeld und die Integration der Lernenden in die Versorgung von Patienten thematisiert. Eine umfassende Übersicht über weitere Instrumente zur Bewertung der Lernumgebung wurde im Jahr 2010 publiziert [16].
Zur Bewertung *individueller Dozenten* stehen zahlreiche Instrumente zur Verfügung (siehe Tabelle 1). Auch hier stehen von den Studierenden ausgefüllte Evaluationsbögen meist mit skalierten Items und Freitextfeldern im Vordergrund. Die medizinspezifischen Dozenten-Evaluationsinstrumente beziehen sich in erster Linie auf die klinische Lehre (z.B. Unterricht am Krankenbett) und weniger auf Vorlesungen und Seminare. Details zu den verfügbaren Instrumenten sind der Tabelle 1 zu entnehmen. Hervorzuheben ist ein Bogen zur Bewertung von

Tabelle 1: Zusammenschau aller identifizierten Instrumente zur Lehrevaluation

Name des Erhebungsinstruments (Literaturzitat)	Item-anzahl	Skalen-anzahl	Bezeichnungen der Skalen	Cronbach's α der Skalen
Dimensionen: „Strukturen" und „Prozesse"				
Medical Student Experience Questionnaire (MedSEQ) [9]	32	5	1: Learning, 2: Teaching and Assessment; 3: Organization and Student Understanding of the Program, 4: Community Interaction and Value, 5: Student Support, Resources	0.63–0.80
Marburger Fragebogen zur Evaluation des Lehrangebots in der Medizin [10]	12		Lehrstruktur, Lehrprozess	0.85–0.87
Dimension: „Prozesse"				
Dundee Ready Education Environment Measure (DREEM) [11]	50	5	1: Students' Perception of Teaching, 2: Students' Perception of Teachers, 3: Students' Academic Self-Perceptions, 4: Students' Perceptions of Atmosphere, 5: Students' Social Self-Perceptions	0.91
Medical Instructional Quality (MedIQ) [15]	25	4	1: The Role of the Preceptor in Facilitating Learning, 2: The Role and Context of the Clinical Environment, 3: Opportunities Available to Learn, 4: Active Involvement by the Learner in the Care of Patients	0.86–0.95
Measuring the School Learning Environment Survey (MSLES) [14]	50	7	1: Breadth of Interest, 2: Student Interaction, 3: Organization, 4: Flexibility, 5: Meaningful Learning Experience, 6: Emotional Climate, 7: Nurturance	0.70–0.85
Learning Environment Questionnaire (LEQ) [13]	65	7	1: Evaluative, 2: Academic Enthusiasm, 3: Goal Direction, 4: Authoritarianism, 5: Breadth of Interest, 6: Student Interaction, 7: Intellectual Maturity	0.74–0.87
Dimension: „Dozent"				
Maastricht Clinical Teaching Questionnaire (MCTQ) [49]	16	5	1: Modeling, 2: Coaching, 3: Exploration, 4: Articulation, 5: Safe Learning Environment	0.83–0.96
Clinical Teaching Assessment Form (CTAF) [50]	9	8	1: Knowledge and Analytical Ability, 2: Clarity and Organization, 3: Enthusiasm and Stimulation, 4: Ability to Establish Rapport, 5: Ability to Involve Student in Learning Experiences, 6: Ability to Give Direction and Feedback, 7: Ability to Demonstrate Clinical Skills and Procedures, 8: Accessibility	nicht publiziert
Student Evaluation of Teaching in Outpatient Clinics (SETOC) [17]	15	5	1: Establishing Learning Milieu, 2: Clinical Teaching, 3: General Teaching, 4: Clinical Competence, 5: Global Rating	0.90–0.98
Stanford Faculty Development Program (SFDP-26) [18]	25	7	1: Learning Climate, 2: Control of Session, 3: Communication of Goals, 4: Understanding and Retention, 5: Evaluation, 6: Feedback, 7: Self-directed Learning	0.82–0.95
SFDP-26 German [19]	25	7	1: Etablieren des Lernklimas, 2: Leitung einer Lehreinheit, 3: Zielkommunikation, 4: Fördern von Verstehen und Behalten, 5: Evaluation, 6: Feedback, 7: Fördern von selbstbestimmten Lernen	0.69–0.92
Mayo Teaching Evaluation Form (MTEF) [51]	28	7	1: Establishing a Positive Learning Climate, 2: Control of the Teaching Session, 3: Communication of Goals, 4: Enhancing Understanding and Retention, 5: Evaluation, 6: Feedback, 7: Self-directed Learning	0.92–0.99

Dozenten, die im Kontext der ambulanten Patientenversorgung unterrichten („Student Evaluation of Teaching in Outpatient Clinics"; SETOC [17]). Zudem soll auf den SFDP-26 („Stanford Faculty Development Program" [18])-Bogen hingewiesen werden, der auch in deutscher Übersetzung verfügbar ist [19]. Dieser ursprünglich an der Mayo Clinic entwickelte Bogen bildet die sieben „Stanford-Kriterien guter Lehre" ab und weist somit eine gute theoretische Fundierung auf.

Das *Ergebnis der Lehre*, d.h. der Lernerfolg der Studierenden spiegelt sich, wie oben ausgeführt, nicht nur im Erwerb von Faktenwissen und praktischen Fertigkeiten sondern auch in der Entwicklung einer professionellen Einstellung wider [7], [8]. Leider konnten keine Instrumente identifiziert werden, die dieses Spektrum erschöpfend abbilden. An einigen deutschen Fakultäten werden die Leistungen der Studierenden im schriftlichen Teil des Zweiten Staatsexamens als Indikatoren der Lehrqualität interpretiert [20]. Allerdings wird in Multiple Choice-Prüfungen vorrangig Faktenwissen thematisiert, und Studierende können durch das Wiedererkennen der richtigen Antwort [21] sowie durch Trainings im Umgang mit MC-Fragen [22] ihr Prüfungsergebnis unabhängig von ihrem Faktenwissen steigern. Ähnlichen Limitationen unterliegt der an einigen deutschen Fakultäten eingesetzte Progress Test. Diese formative, wiederholt während des Studiums durchgeführte Prüfungsform verwendet ebenfalls MC-Fragen. Der Test wird durch das quer- und längsschnittliche Design jedoch als sinnvolle Quelle für wichtige Informationen für den Lernenden sowie für die Curriculumsevaluation angesehen [23].

Die staatliche Examensprüfung weist in der Regel eine hohe interne Konsistenz auf. Zur Bewertung des Ergebnisses einzelner Veranstaltungen innerhalb einer Fakultät müssen jedoch fakultätsinterne Prüfungen herangezogen werden. Diese genügen einer aktuellen Analyse zufolge oft nicht den Qualitätsstandards [24]. Als Alternative zur Messung des Lehr-Ergebnisses anhand von Prüfungsleistungen wurde kürzlich ein Instrument zur Abschätzung des Lernerfolgs anhand wiederholter studentischer Selbsteinschätzungen entwickelt. Dieses bietet gegenüber Abschlussprüfungen den Vorteil, dass es auch den initialen Leistungsstand der Studierenden berücksichtigt und somit Aussagen über den tatsächlichen Lernzuwachs während einer Veranstaltung zulässt [25].

Schließlich ist als Methode zur Bewertung der medizinischen Lehre die Absolventenbefragung zu nennen. Prinzipiell können mit dieser Methode alle vier Dimensionen der Lehrqualität betrachtet werden. Im Rahmen der vorliegenden Literatursuche wurden jedoch weder medizinspezifische Forschungsarbeiten noch Studien aus anderen Bereichen der Hochschullehre identifiziert, in denen die Qualität der hierzu genutzten Instrumente systematisch untersucht wurde.

Frage 3: Psychometrische Eigenschaften der Erhebungsinstrumente

Sowohl studentische Evaluationsbögen als auch Prüfungen können hinsichtlich ihrer Reliabilität und Validität beurteilt werden. Die *Reliabilität* der Instrumente zur Bewertung lehrbezogener Strukturen und Prozesse ist der letzten Spalte von Tabelle 1 zu entnehmen. Das Cronbach's α als unteres Grenzmaß der Reliabilität ist für die meisten betrachteten Fragebogen-Instrumente zufriedenstellend. Die Interrater-Reliabilität der Evaluation hängt von der Anzahl der ausgefüllten Evaluationsbögen ab [26]. Allerdings liegen bislang keine Studien dazu vor, welcher absolute Rücklauf mindestens erforderlich ist, um aussagekräftige Daten zu erhalten (s.u.). Die Messung der Reliabilität fakultätsinterner Prüfungen stellt eine wesentliche Voraussetzung für deren Nutzung zu Evaluationszwecken dar. Bislang werden entsprechende statistische Analysen allerdings für weniger als 40% der Leistungsnachweise an deutschen Medizinischen Fakultäten angestellt [24].

Eine inhaltlich fundierte Interpretation von Evaluationsergebnissen setzt voraus, dass die erhobenen Daten valide sind. Während viele Evaluationsbögen und Prüfungen eine akzeptable Inhaltsvalidität aufweisen, sind in der Regel keine Informationen über ihre Kriteriums- und Konstruktvalidität verfügbar. Zu berücksichtigen sind außerdem Störfaktoren, die sich auf die Validität der Ergebnisse auswirken können und in erster Linie bei *studentischen Bewertungen* identifiziert wurden. Diese möglichen Faktoren sind unten genauer dargestellt (Frage 4); allerdings stützen sich die Betrachtungen zu diesem Aspekt vorrangig auf Literatur ohne direkten Bezug zum Medizinstudium.

Die Validität von *Prüfungen* wird im Wesentlichen durch zwei Störfaktoren gefährdet [27]. Eine „Konstrukt-Unterrepräsentation" liegt dann vor, wenn das zu prüfende Konstrukt in der Prüfung nicht erschöpfend behandelt wird. In diesem Fall sind Studierende im Vorteil, die (zufällig) diejenigen Inhalte intensiver gelernt haben, die von der Prüfung abgedeckt wurden. Der zweite wesentliche Störfaktor ist die „Konstrukt-irrelevante Varianz"; sie entsteht beispielsweise dann, wenn Prüfungsfragen suboptimal konstruiert sind, so dass nicht nur die offensichtlichen Inhalte, sondern auch die Befähigung der Studierenden zum Umgang mit schwer verständlichen Formulierungen geprüft wird. Die Kriteriumsvalidität von Prüfungen wird im praktischen Lehr-Alltag in Ermangelung eines validen Außenkriteriums und der erforderlichen Ressourcen meist nicht überprüft. Das oben erwähnte Instrument zur Abschätzung des studentischen Lernerfolgs anhand wiederholter Selbsteinschätzungen hat sich in einer ersten Studie als konstruktvalide erwiesen [25]; weitere publizierte Ergebnisse lagen zum Zeitpunkt der hier vorgestellten Literatursuche noch nicht vor. Auch konnten keine Studien zur Reliabilität und Validität von Absolventenbefragungen identifiziert werden.

Frage 4: Fragebogenkonstruktion und Störgrößen

Das dominierende Erhebungsinstrument sowohl in der Praxis als auch in den identifizierten Publikationen sind von Studierenden auszufüllende Fragebögen. Bei der Konstruktion und dem Einsatz von Fragbögen sind jedoch einige Aspekte zu berücksichtigen. Wie oben bereits angemerkt, ließ sich für diese Forschungsfrage bedauerlicherweise kaum medizinspezifische Literatur identifizieren. Somit sollen im Folgenden einige einschlägige Erkenntnisse zur Fragebogenkonstruktion und zu den wichtigsten Störgrößen studentischer Lehrevaluationsinstrumente vornehmlich ohne direkten Bezug zum Medizinstudium dargestellt werden.

Sowohl das Fragenformat als auch die Antwortskala und das Erhebungsformat können sich auf die psychometrischen Eigenschaften der Instrumente auswirken. Bezüglich des Fragenformats wird zunächst zwischen Freitextfragen und skalierten Items unterschieden. Frei formulierte Evaluationskommentare können wertvolle qualitative Informationen liefern, werden aber nicht von allen Studierenden abgegeben. Skalierte Items bilden die Grundlage quantitativer Analysen. Die häufig verwendeten globalen Items zur Gesamtbewertung einer Veranstaltung (z.B. nach dem Schulnotenprinzip) werden von einigen Autoren aufgrund ihrer Anfälligkeit für verzerrende Einflüsse (s.u.) kritisiert [28], [29]. Andere Autoren vertreten hingegen die Ansicht, dass gute Lehre als nahezu eindimensionales Konstrukt gut mittels globaler Items beurteilt werden kann [30]. Wieder andere Studien zeigen, dass die Reliabilität eines Instruments umso höher ist, je mehr spezifische Items es enthält [31], [32].

Bezüglich der Skalierung der Antwortoptionen ist anzumerken, dass generell bessere Bewertungen zu erwarten sind, wenn sich der positive Anker der Skala links befindet [33]. Des Weiteren ist bekannt, dass die Formulierungen der Items nicht von allen Studierenden gleich interpretiert werden [3]. Auch das Erhebungsformat ist zu berücksichtigen. Diesem kommt eine wachsende Bedeutung zu, da viele Fakultäten die Evaluation ihrer Lehrveranstaltungen mittlerweile über Online-Plattformen abwickeln. Online-Evaluationen gehen zumeist mit einem geringeren Rücklauf einher als traditionelle Papier-Evaluationen. Wenngleich sich dies einer Studie zufolge nicht auf die Evaluationsergebnisse auswirkt und online sogar mehr Freitext-Kommentare abgegeben werden [34], wurde auch berichtet, dass sich leistungsschwache Studierende weniger an Online-Evaluationen beteiligen als leistungsstarke [35]. Des Weiteren ist anzumerken, dass anonyme Befragungen in der Regel schlechtere Bewertungen liefern als Befragungen, in denen die Studierenden sich identifizieren müssen [36]. Hinsichtlich der oben erwähnten Absolventenbefragungen ist zu beachten, dass Evaluationsergebnisse umso schlechter ausfallen, je größer der zeitliche Abstand zwischen der Lehre und ihrer Bewertung ist [37].

Items, die zur Bewertung individueller Dozenten eingesetzt werden, zeigen sich besonders störanfällig. So wurde wiederholt gezeigt, dass Lehrende, die enthusiastisch auftreten oder eine gute Reputation [38] haben, systematisch besser bewertet werden – selbst wenn die von ihnen vermittelten Inhalte fehlerhaft sind [39], [40]. Das studentische Interesse an einem Kurs ist ein weiterer wichtiger Störfaktor [41], [42] – folglich werden Wahlkurse in der Regel besser bewertet als Pflichtkurse [28], [43]. Veranstaltungen, die besser besucht sind, erhalten ebenfalls zumeist positivere Bewertungen [44]. Speziell im Medizinstudium werden die Veranstaltungen theoretischer Fächer tendenziell schlechter bewertet als die Lehre in klinischen Fächern; ebenso erhalten Vorlesungen im Schnitt schlechtere Bewertungen als Kleingruppenunterricht [37].

Diskussion

Die vorliegende Arbeit ist das Ergebnis einer breit angelegten Bestandsaufnahme der verfügbaren Literatur zur Evaluation der Lehre in der Humanmedizin. Die Ergebnisse der Recherche unterstreichen nochmals, dass die Qualität der Lehre kein eindimensionales Konstrukt darstellt; vielmehr können und sollten in Evaluationen alle vier – teilweise überlappenden – Dimensionen „Struktur“, „Prozess“, „Dozent“ und „Ergebnis“ betrachtet werden. Außerdem muss die Beurteilung von Evaluationsergebnissen stets vor dem Hintergrund des Konstrukts erfolgen, das dem genutzten Instrument zugrunde liegt. Konkret können aus studentischen Bewertungen der Pünktlichkeit von Dozenten oder der räumlichen Gegebenheiten an einer Hochschule keine unmittelbaren Rückschlüsse auf den Lernerfolg der Studierenden gezogen werden. Prüfungsergebnisse können zwar zur Abschätzung des Lehr-Ergebnisses herangezogen werden; sie bilden jedoch in der Regel nur den Leistungsstand zu einem festen Zeitpunkt ab und erlauben keine Bewertung des Lernerfolgs im Laufe einer Lehrveranstaltung. Der Progress Test überwindet diese Einschränkung zwar durch seine wiederholte Durchführung, beinhaltet jedoch nur MC-Fragen und bildet daher keine praktischen Fertigkeiten oder professionellen Einstellungen ab. Außerdem wird er nicht vor und nach jedem Kurs/Modul durchgeführt – dies wäre aber erforderlich, um einzelne Kurse/Module (und nicht nur Studierendenkohorten bzw. ganze Studiengänge) zu evaluieren.

Die quantitative Analyse von Evaluationsdaten (z.B. durch Mittelwertbildung der studentischen Globalbewertung eines Kurses auf einer Schulnoten-Skala) eröffnet zwar die Möglichkeit des Vergleichs zwischen Veranstaltungen; dieses Vorgehen birgt aber zwei Risiken: Erstens wird mit Globalbewertungen wahrscheinlich ein nicht trennscharf definiertes Konstrukt abgebildet, und zweitens sind solche Bewertungen einer Vielzahl verzerrender Einflüsse unterworfen [45]. Beides wirkt sich mindernd auf die Reliabilität und Validität von Globalbewertungen aus, falls angenommen wird, dass die Qualität der gesamten Lehre mit

allen Facetten durch eine einzige Kennzahl abgebildet werden kann. Zusätzlich zu den zahlreichen oben genannten Störfaktoren soll an dieser Stelle auch der Umfang der Erhebungsinstrumente erwähnt werden. Einige der in Tabelle 1 aufgeführten Bögen enthalten über 60 Items und eignen sich aufgrund mangelnder studentischer Akzeptanz wahrscheinlich nicht zum Einsatz im Rahmen einer regelmäßigen und häufigen Veranstaltungsevaluation [46].

Weniger als die Hälfte der identifizierten Volltext-Arbeiten haben Eingang in die hier präsentierte Zusammenstellung gefunden. Hauptgrund für den Ausschluss der meisten Arbeiten war ihr fehlender Bezug zum Medizinstudium. Sehr weit verbreitet ist im amerikanischen Raum beispielsweise der validierte Fragenbogen SEEQ („Students' Evaluation of Educational Quality") [47]. Ob dieses Instrument auf die Medizin übertragbar ist, ist fraglich: Zum einen ist es für die amerikanische Hochschullehre entwickelt worden, die nur eingeschränkt mit der deutschen vergleichbar ist, zum anderen ist es kein medizinspezifisches Instrument. Weithin bekannte deutschsprachige Instrumente zur Evaluation der (nicht-medizinischen) Hochschullehre sind HILVE („Heidelberger Inventar zur Lehrveranstaltungs-Evaluation") [48] und HILVE II. Beide besitzen gute psychometrische Charakteristika, aber auch hier stellt sich die Frage nach der Übertragbarkeit auf den medizinischen Kontext. Aufgrund der eingangs genannten Besonderheiten des Medizinstudiums erscheint auf jeden Fall vor einem entsprechenden Einsatz eine erneute psychometrische Testung in diesem Setting geboten.

Generelle Empfehlungen zum Einsatz spezifischer Instrumente im Medizinstudium an deutschen Fakultäten lassen sich aus den Ergebnissen der Literatursuche nicht ableiten, da die Wahl des Instruments sich wie oben dargestellt am Evaluationsziel orientieren sollte. Eine vorläufige, ressourcensparende Lösung könnte darin bestehen, die bereits in deutscher Sprache verfügbaren und mit guten psychometrischen Charakteristika ausgestatteten Bögen Marburger Fragebogen (für Strukturen und Prozesse) und SFDP-26 German [19] (für Dozenten) einzusetzen. Aufgrund der eingeschränkten Übertragbarkeit der vornehmlich im angelsächsischen Sprachraum entwickelten und validierten Instrumente auf den Kontext des deutschen Medizinstudiums sollte mittelfristig angestrebt werden, aus bereits verfügbaren, teilweise aber auch neu konstruierten Items einen neuen Fragebogen zu erstellen, der dann direkt an deutschen Fakultäten evaluiert wird. Dieser Prozess muss von psychometrischer Expertise begleitet werden und könnte im Rahmen eines entsprechenden Forschungsprojekts mehrere interessierte medizinische Fakultäten einbeziehen. Mit Hilfe eines gemeinsam konsentierten Instrumentes könnte durch die Nutzung an mehreren Standorten eine höhere Vergleichbarkeit der Ergebnisse erreicht werden. Bezüglich einer möglichen Entwicklungs- und Implementierungsstrategie finden zurzeit weitergehende Konsultationen zwischen MFT und AWMF statt.

Neben der Möglichkeit, dass relevante Publikationen in unserer Literatursuche nicht enthalten sind, ist die Hauptlimitation der vorliegenden Arbeit, dass ein Großteil der betrachteten Literatur aus dem anglo-amerikanischen Sprachraum stammt mit zuweilen erheblichen Unterschieden gegenüber dem Medizinstudium in Deutschland (z.B. eingeschränkte Übertragbarkeit angelsächsischer „clerkships" auf deutsche Blockpraktika und Famulaturen; Fehlen eines direkten Äquivalents zum Praktischen Jahr in den angelsächsischen Studiengängen). Zudem bezogen sich die Quellen, auf die bei der vierten Forschungsfrage zurückgegriffen wurde, größtenteils nicht primär auf das Medizinstudium. Es ist zumindest fraglich, ob Erkenntnisse zur Fragebogenkonstruktion und Störgrößen in der Evaluation aus anderen Disziplinen sich ohne weiteres auf das Studium der Humanmedizin übertragen lassen. Schließlich lieferte die von uns angestellte Suche nach publizierten Instrumenten zur Beurteilung der Lehrqualität hauptsächlich Fragebögen, die im Rahmen einer studentischen Evaluation eingesetzt werden können. Andere Verfahren (z.B. Absolventenbefragungen) könnten ebenfalls hilfreiche Informationen liefern; aufgrund der diesbezüglich limitierten Datenlage wurde auf eine entsprechende Diskussion im Rahmen dieser Übersicht verzichtet.

Fazit

Die Evaluation der medizinischen Hochschullehre stützt sich in erster Linie auf studentische Bewertungen, die sich auf lehrbezogene Strukturen und Prozesse sowie die Leistung individueller Dozenten beziehen. In der vorliegenden Recherche wurden einige reliable Instrumente zur Betrachtung dieser drei Dimensionen der Lehrqualität identifiziert; allerdings sind zumindest einige Störfaktoren aus nicht medizinichscher Literatur bekannt, die sich auf das studentische Bewertungsverhalten auswirken und somit die Validität der Erhebungsinstrumente einschränken. Diese Störfaktoren sollten auch bei der Nutzung studentischer Evaluationen zur Bewertung der medizinischen Lehre Berücksichtigung finden bzw. neu geprüft werden. Die Bewertung der Lehrqualität anhand von Prüfungsergebnissen ist aufgrund der bisher ungesicherten Qualität fakultätsinterner Prüfungen in Deutschland problematisch; Absolventenbefragungen werden nicht flächendeckend und mit Instrumenten ungewisser Validität und Reliabilität durchgeführt.

Konsequenzen für Klinik und Praxis

- Die Qualität der medizinischen Lehre ist ein mehrdimensionales Konstrukt; die wesentlichen vier Dimensionen, anhand derer die Lehrqualität beurteilt werden kann, sind Strukturen, Prozesse, Dozenten-Charakteristika und das Lehr-Ergebnis.
- Für die Bewertung von Strukturen, Prozessen und individuellen Dozenten im Medizinstudium stehen verschiedene Instrumente mit guten psychometrischen Charakteristika zur Verfügung. Die Messung des Lehr-

Ergebnisses ist aufgrund der größtenteils unbekannten bzw. unbefriedigenden Reliabilität und Validität fakultätsinterner Prüfungen zurzeit noch erheblichen Limitationen unterworfen.

- Bei der Konzeption und Nutzung von Evaluationsinstrumenten müssen die in dieser Arbeit dargestellten Störgrößen berücksichtigt werden, insofern diese aus anderen Lehr-Kontexten bekannten Faktoren auf das Medizinstudium übertragbar sind.

Anmerkungen

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Autorenschaft

Die Autoren Herrmann-Lingen C and Raupach T haben gleichermaßen zu der Arbeit beigetragen.

Anhänge

Verfügbar unter
http://www.egms.de/en/journals/gms/2015-13/000219.shtml

1. 000219_Appendix.pdf (151 KB)
   Aufstellung der Gesamtliteratur

Literatur

1. DeGEval – Gesellschaft für Evaluation e.V., editor. Standards für Evaluation. Köln: DeGEval; 2002.
2. Kogan JR, Shea JA. Course evaluation in medical education. Teach Teach Educ. 2007;23(3):251-64. DOI: 10.1016/j.tate.2006.12.020
3. Billings-Gagliardi S, Barrett SV, Mazor KM. Interpreting course evaluation results: insights from thinkaloud interviews with medical students. Med Educ. 2004 Oct;38(10):1061-70. DOI: 10.1111/j.1365-2929.2004.01953.x
4. Colliver JA, Kucera K, Verhulst SJ. Meta-analysis of quasi-experimental research: are systematic narrative reviews indicated? Med Educ. 2008 Sep;42(9):858-65. DOI: 10.1111/j.1365-2923.2008.03144.x
5. Eva KW. On the limits of systematicity. Med Educ. 2008 Sep;42(9):852-3. DOI: 10.1111/j.1365-2923.2008.03140.x
6. Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP. Enhancing evaluation in an undergraduate medical education program. Acad Med. 2008 Aug;83(8):787-93. DOI: 10.1097/ACM.0b013e31817eb8ab
7. Blumberg P. Multidimensional outcome considerations in assessing the efficacy of medical educational programs. Teach Learn Med. 2003;15(3):210-4. DOI: 10.1207/S15328015TLM1503_10
8. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach. 2007 Sep;29(7):642-7. DOI: 10.1080/01421590701746983
9. Boyle P, Grimm MC, McNeil HP, Scicluna H. The UNSW Medicine Student Experience Questionnaire (MedSEQ). San Francisco: Academia; 2009. Available from: http://www.academia.edu/5252480/Medicine_Student_Experience_Questionnaire_MEDSEQ_UNSW
10. Krebs K. Marburger Fragebogen zur Evaluation des Lehrangebots in der Medizin: Eine Untersuchung zur Reliabilität und Dimensionalität des Marburger Fragebogens zur Evaluation des Lehrangebots am Fachbereich Medizin [Dissertation]. Marburg: Philipps-Universität Marburg; 2006. Available from: http://archiv.ub.uni-marburg.de/diss/z2006/0387/pdf/dkk.pdf
11. Roff S. The Dundee Ready Educational Environment Measure (DREEM)– a generic instrument for measuring students' perceptions of undergraduate health professions curricula. Med Teach. 2005 Jun;27(4):322-5. DOI: 10.1080/01421590500151054
12. Rotthoff T, Ostapczuk MS, De Bruin J, Decking U, Schneider M, Ritz-Timme S. Assessing the learning environment of a faculty: psychometric validation of the German version of the Dundee Ready Education Environment Measure with students and teachers. Med Teach. 2011;33(11):e624-36. DOI: 10.3109/0142159X.2011.610841
13. Rothman AI, Ayoade F. The development of a learning environment: a questionnaire for use in curriculum evaluation. J Med Educ. 1970;45(10):754-9. DOI: 10.1097/00001888-197010000-00006
14. Marshall RE. Measuring the medical school learning environment. Acad Med. 1978;53(2):98-104. DOI: 10.1097/00001888-197802000-00003
15. James PA, Osborne JW. A measure of medical instructional quality in ambulatory settings: the MedIQ. Fam Med. 1999 Apr;31(4):263-9.
16. Soemantri D, Herrera C, Riquelme A. Measuring the educational environment in health professions studies: a systematic review. Med Teach. 2010;32(12):947-52. DOI: 10.3109/01421591003686229
17. 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 Feb;12(1):55-69. DOI: 10.1007/s10459-005-2328-y
18. 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):688-95. DOI: 10.1097/00001888-199806000-00016
19. Iblher P, Zupanic M, Härtel C, Heinze H, Schmucker P, Fischer MR. Der Fragebogen "SFDP26-German": Ein verlässliches Instrument zur Evaluation des klinischen Unterrichts? [The Questionnaire "SFDP26-German": a reliable tool for evaluation of clinical teaching?]. GMS Z Med Ausbild. 2011;28(2):Doc30. DOI: 10.3205/zma000742
20. Herzig S, Marschall B, Nast-Kolb D, Soboll S, Rump LC, Hilgers RD. Positionspapier der nordrhein-westfälischen Studiendekane zur hochschulvergleichenden leistungsorientierten Mittelvergabe für die Lehre [Distribution of government funds according to teaching performance]. GMS Z Med Ausbild. 2007;24(2):Doc109. Available from: http://www.egms.de/en/journals/zma/2007-24/zma000403.shtml
21. Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen [Format and implementation of written assessments]. GMS Z Med Ausbild. 2006; 23(3):Doc44. Available from: http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
22. Mahamed A, Gregory PA, Austin Z. "Testwiseness" among international pharmacy graduates and Canadian senior pharmacy students. Am J Pharm Educ. 2006 Dec;70(6):131. DOI: 10.5688/aj7006131

23. Freeman A, Van Der Vleuten C, Nouns Z, Ricketts C. Progress testing internationally. Med Teach. 2010;32(6):451-5. DOI: 10.3109/0142159X.2010.485231

24. Möltner A, Duelli R, Resch F, Schultz JH, Jünger J. Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten [School-specific assessment in German medical schools]. GMS Z Med Ausbild. 2010;27(3):Doc44. DOI: 10.3205/zma000681

25. 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):e446-53. DOI: 10.3109/0142159X.2011.586751

26. Spiel C, Schober B, Reimann R. Evaluation of curricula in higher education: challenges for evaluators. Eval Rev. 2006 Aug;30(4):430-50. DOI: 10.1177/0193841X05285077

27. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327-33. DOI: 10.1046/j.1365-2923.2004.01777.x

28. Aleamoni LM. Student rating myths versus research facts from 1924 to 1998. J Pers Eval Educ. 1999;13(2):153-66. DOI: 10.1023/A:1008168421283

29. Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. Am Psychol. 1997;52(11):1187-97. DOI: 10.1037/0003-066X.52.11.1187

30. d'Apollonia S, Abrami PC. Navigating student ratings of instruction. Am Psychol. 1997;52(11):1198-208. DOI: 10.1037/0003-066X.52.11.1198

31. Jackson DL, Teal CR, Raines SJ, Nansel TR, Force RC, Burdsal CA. The dimensions of students' perceptions of teaching effectiveness. Educ Psychol Meas. 1999;59(4):580-96. DOI: 10.1177/00131649921970035

32. Marsh HW. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. J Educ Psychol. 1984;76(5):707-54. DOI: 10.1037/0022-0663.76.5.707

33. 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 Jul;72(7):627-30. DOI: 10.1097/00001888-199707000-00015

34. Sorenson DL, Johnson TD. Online student ratings of instruction. New Dir Teach Learn. 2003;2003(96):1-112.

35. 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):576-91. DOI: 10.1007/s11162-011-9240-5

36. Afonso NM, Cardozo LJ, Mascarenhas OA, Aranha AN, Shah C. Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. Fam Med. 2005 Jan;37(1):43-7.

37. van den Bussche H, Weidtmann K, Kohler N, Frost M, Kaduszkiewicz H. Evaluation der ärztlichen Ausbildung: Methodische Probleme der Durchführung und der Interpretation von Ergebnissen [Evaluation of medical education: methodological problems of implementation and interpretation of results]. GMS Z Med Ausbild. 2006;23(2):Doc37. Available from: http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml

38. Griffin BW. Instructor reputation and student ratings of instruction. Contemp Educ Psychol. 2001 Oct;26(4):534-52. DOI: 10.1006/ceps.2000.1075

39. Marsh HW, Ware JE. Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. J Educ Psychol. 1982;74(1):126-34. DOI: 10.1037/0022-0663.74.1.126

40. Naftulin DH, Ware JE, Donnelly FA. The Doctor Fox Lecture: a paradigm of educational seduction. J Med Educ. 1973 Jul;48(7):630-5. DOI: 10.1097/00001888-197307000-00003

41. Prave RS, Baril GL. Instructor ratings: Controlling for bias from Initial student interest. J Educ Bus. 1993;68(6):362-6. DOI: 10.1080/08832323.1993.10117644

42. Cashin WE. Student ratings of teaching: A summary of the research. East Lansing, MI, USA: Office of Faculty and Organizational Development at Michigan State University; 1988. (IDEA Paper; No.20). Available from: http://ideaedu.org/wp-content/uploads/2014/11/idea-paper_50.pdf

43. Ting KF. A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. Res High Educ. 2000;41(5):637-61. DOI: 10.1023/A:1007075516271

44. Abrami PC, D'Apollonia S, Cohen PA. Validity of student ratings of instruction: What we know and what we do not. J Educ Psychol. 1990;82(2):219-31. DOI: 10.1037/0022-0663.82.2.219

45. Schiekirka S, Raupach T. A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. BMC Med Educ. 2015 Mar 5;15:30. DOI: 10.1186/s12909-015-0311-8

46. Schiekirka S, Reinhardt D, Heim S, Fabry G, Pukrop T, Anders S, Raupach T. Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. BMC Med Educ. 2012 Jun 22;12:45. DOI: 10.1186/1472-6920-12-45

47. Marsh HW. SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. Brit J Psychol. 1982;52(1):77-95. DOI: 10.1111/j.2044-8279.1982.tb02505.x

48. Rindermann H, Schofield N. Generalizability of Multidimensional Student Ratings of University Instruction Across Courses and Teachers. Res High Educ. 2001;42(4):377-99. DOI: 10.1023/A:1011050724796

49. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. Acad Med. 2010 Nov;85(11):1732-8. DOI: 10.1097/ACM.0b013e3181f554d6

50. Irby DM, Gillmore GM, Ramsey PG. Factors affecting ratings of clinical teachers by medical students and residents. Acad Med. 1987;62(1):1-7. DOI: 10.1097/00001888-198701000-00001

51. Beckman TJ, Lee MC, Rohren CH, Pankratz VS. Evaluating an instrument for the peer review of inpatient teaching. Med Teach. 2003 Mar;25(2):131-5. DOI: 10.1080/0142159031000092508

Korrespondenzadresse:

Dipl.-Psych. Sarah Schiekirka
Universitätsmedizin Göttingen, Studiendekanat, Humboldtallee 38, 37073 Göttingen, Deutschland, Tel.: +49-(0)551-39-12302, Fax: +49-(0)551/39-13012302
sarah.schiekirka@med.uni-goettingen.de

Bitte zitieren als
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. GMS Ger Med Sci. 2015;13:Doc15.
DOI: 10.3205/000219, URN: urn:nbn:de:0183-0002197

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/gms/2015-13/000219.shtml

Eingereicht: *02.04.2015*
Überarbeitet: *31.08.2015*
Veröffentlicht: *16.09.2015*