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

• Hendrik van den Bussche[1] • Katja Weidtmann[2] • Nikolaj Kohler[1] • Maike Frost[2] • Hanna Kaduszkiewicz[1]

Zusammenfassung:
Hintergrund:
Die Evaluation von Lehrveranstaltungen und curricularen Abschnitten durch Studierende nimmt nicht nur in der ärztlichen Ausbildung zu. Neben der Rückmeldung an die Dozenten können die Ergebnisse eine wichtige Basis für budgetäre Allokationsentscheidungen in der Lehre bilden. An der Medizinischen Fakultät der Universität Hamburg wurden seit zehn Jahren Erfahrungen mit der Evaluation von Lehrveranstaltungen gesammelt.
Ergebnisse:
Eine Analyse der Hamburger Erfahrungen zeigt, dass die Ergebnisse standardisierter Evaluationen der ärztlichen Ausbildung eine Reihe von Problemen in sich bergen:
1. Die Gesamtbeurteilung eines Studienabschnitts scheint anders auszufallen als die Beurteilung der darin enthaltenen einzelnen Veranstaltungen.
2. Theoretische bzw. patientenferne Fächer werden - in der Regel unabhängig von deren tatsächlicher Qualität - insgesamt negativer beurteilt als klinische Fächer. Grund dafür ist wahrscheinlich der als geringer eingeschätzte Nutzen der patientenfernen Fächer für die klinische Berufstätigkeit.
3. Es zeigen sich Unterschiede zwischen zeitnahen und zeitfernen Veranstaltungsevaluationen.
4. Die Bewertung durch Studierende ist weiterhin abhängig von deren Studiendauer sowie von der Veranstaltungsform.
Schlussfolgerung:
Beim Ranking von Fächern und Veranstaltungen ist Vorsicht geboten. Zur Verbesserung der Interpretationssicherheit können folgende Empfehlungen gegeben werden:
1. Aus Gesamtbeurteilungen sollte nur mit Vorsicht auf die einzelnen Veranstaltungen geschlossen werden.
2. Beim Ranking sind "Handicaps" der theoretischen bzw. patientenfernen Fächer zu bedenken.
3. Mehrfachmessungen und Längsschnittvergleiche sind unerlässlich.
4. Vergleiche sollten nur bei vergleichbaren Einheiten (z.B. Veranstaltungstypen) vorgenommen werden.
5. Empirisch gefundene Unterschiede sind immer auf ihre Ursachenbündel zu untersuchen, am besten von gemischten Evaluationskommissionen.
6. Erhebungen zur Prozessqualität von Veranstaltungen sind brauchbarer als solche zur Ergebnisqualität.
Schlüsselwörter: Evaluation, Methodik, medizinische Ausbildung

Abstract:
Background:
The importance of evaluation is increasing, not only in medical education. Apart from giving feedback to the teachers the results of evaluations can be used for allocation of educational funds. At the medical faculty of the University of Hamburg the evaluation of medical education has been performed since ten years.
Results:
An analysis of the Hamburg experience in the standardized evaluation of medical education shows a series of problems.
1. The cumulative evaluation of curriculum parts tends to turn out different than the detailed evaluation of its consisting parts.
2. Theoretical courses tend to be assessed worse than clinical courses - irrespective of their real quality. The reason for this "handicap" of theoretical courses is that students tend to rate their benefit for clinical practice lower.
3. There are differences between prompt and delayed assessments of the same courses.
4. Students' assessment also depends on the stage of their training and on the didactic type of the evaluated course.
Conclusion:
Ranking of courses should be performed with caution. To improve reliability and validity of results the following suggestions are made:
1. Results of cumulative evaluation of curriculum parts should be interpreted with care.
2. Unequal opportunities of theoretical and clinical courses to gain good results should be considered.
3. Repeated assessments and longitudinal comparisons are essential.
4. Comparisons should be made only between similar didactical types of courses.

[1] *Universitätsklinikum Hamburg-Eppendorf, Institut für Allgemeinmedizin, Zentrum für Psychosoziale Medizin, Hamburg, Deutschland*

[2] *Universitätsklinikum Hamburg-Eppendorf, Prodekanat für Lehre, Hamburg, Deutschland*

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.

5. Results should be analysed with respect to their possible causes, preferably by mixed evaluation commissions.
6. It is recommended to evaluate aspects of process quality rather than of result quality.
Keywords: evaluation, methodology, medical education

Einleitung

Die Evaluation von Lehrveranstaltungen und curricularen Abschnitten nimmt (nicht nur) in der ärztlichen Ausbildung zu [1][2][3]. Neben der didaktischen Funktion können die Ergebnisse der Evaluation von Ausbildungsangeboten eine wichtige Basis für budgetäre Allokationsentscheidungen in der Lehre bilden. Daraus ergibt sich die berechtigte Forderung nach einer angemessenen Reliabilität und Validität von Evaluationsergebnissen. Dass dies leichter gefordert als eingelöst ist, soll die folgende Analyse von Problemen der Evaluation der ärztlichen Ausbildung verdeutlichen. Sie beruht auf zehnjähriger Erfahrung mit der Evaluation von Lehrveranstaltungen an der Hamburger Medizinischen Fakultät.

Eine Problemauflistung

Es ist eine Binsenweisheit, dass die Ergebnisse von Erhebungen durch die angewandte Erhebungsmethode und durch die Erhebungssituation beeinflusst werden. Dennoch wird in den meisten Erhebungen nur eine Methode benutzt oder die Erhebung auf eine einmalige Messung beschränkt. In der Evaluation der ärztlichen Ausbildung wird vielfach - so auch an der Hamburger Medizinischen Fakultät - in erster Linie auf standardisierte Fragebögen zurückgegriffen, in denen die Studierenden unter Wahrung ihrer Anonymität likertskalierte Statementfragen bewerten (vgl. Abbildung 1).

Aussagen zu den Seminaren	trifft gar nicht zu					sehr zu
Die Seminare waren gut strukturiert	0	0	0	0	0	0
Ich habe durch die Seminare viel dazugelernt	0	0	0	0	0	0
entspricht Werten von	1	2	3	4	5	6

Abbildung 1: Beispiel aus den Erhebungsinstrumenten

Der Vorteil dieser Methode liegt in ihrer geringen Personalintensität und damit ihrer Kostengünstigkeit, ferner in der Gleichbehandlung aller untersuchten Einheiten, seien es Veranstaltungen oder wissenschaftliche Einrichtungen. Dies verleiht ihnen einen Anschein von Objektivität. Diese "einheitliche" Methodik geht aber mit einer Reihe von Problemen einher. In der wissenschaftlichen Literatur finden sich viele Arbeiten zu Stör- bzw. Biasvariablen der studentischen Lehrevaluation. Als von vielen Seiten akzeptierte, wenn auch nicht unumstrittene, Biasvariablen gelten das vorbestehende Interesse am Fach bzw. an der Veranstaltung, die Notenerwartung, die Arbeitslast bzw. Schwierigkeit der Veranstaltung und die Gruppengröße [4][5][6][7]. In der zehnjährigen Evaluationstätigkeit der Lehre an der Hamburger Medizinischen Fakultät sind weitere Probleme mehrfach aufgetreten und weisen somit einen gewissen überzufälligen Charakter auf. Diese Probleme werden in diesem Artikel dargestellt und anhand von Beispielen verdeutlicht. Die nachfolgende Beschreibung beansprucht aber keinen Anspruch auf Vollständigkeit und ist kein Beweis dafür, dass solche Probleme immer auftreten müssen. Ziel dieser Publikation ist auf die Gefahr möglicher Fehlinterpretationen hinzuweisen und den Blick dafür zu schärfen, dass die Beurteilung von Evaluationsdaten mit Vorsicht vorzunehmen ist.

Eine methodische Vorbemerkung: Streng messtheoretisch sind die in den als Beispiele angeführten Erhebungen verwendeten Ratingskalen ordinalskaliert, was die Interpretation der Skalenmittelwerte problematisch macht [8]. Aus didaktischen Gründen wurde die Darstellung auf der Basis von Mittelwerten dennoch beibehalten.

• 1. Unterschiede zwischen der Bewertung von Einzelfächern im Vergleich zu gesamten Curriculumsabschnitten

Mehrfach wurde festgestellt, dass die Bewertung eines gesamten Studienabschnitts sich von der Bewertung der darin enthaltenen einzelnen Fächer bzw. Veranstaltungen beträchtlich unterscheiden kann. Hierzu folgende Beispiele:

1.1. Die vorklinische Ausbildung

In einer retrospektiven Evaluation der vorklinischen Ausbildung im Jahr 2001 [9] wurde festgestellt, dass die Bewertung der Vorklinik als ganzen Studienabschnitt überwiegend deutlich kritischer ausfiel als die Bewertung der Mehrzahl der einzelnen Fächer. So gaben zwei Drittel (65%) der im 5. klinischen Semester retrospektiv befragten 170 Studierenden bei einer Rücklaufquote von 88% an, mit der Vorklinik "insgesamt nicht zufrieden" zu sein (M=2.2 auf einer Vierpunktskala, SD=0.98). Mehr als die Hälfte der Studierenden war der Meinung, "das in der Vorklinik Gelernte sei (eher) nicht von großem Nutzen für das weitere klinische Studium." (M=2.5; SD=0.85; vgl. Abbildung 2).

Abbildung 2: Antworthäufigkeiten zum Nutzen der vorklinischen Ausbildung (M=2.5; SD=0,85)

Demgegenüber ergab die Frage nach dem Nutzen der einzelnen Fächer bzw. Veranstaltungen ein anderes Bild: Immerhin wurden z.B. die Veranstaltungen des Faches Anatomie von den gleichen Studierenden in der gleichen Befragung als eher nützlich beurteilt (Werte zwischen drei und vier in der nachfolgenden Abbildung 3).

Nennenswerte Unterschiede nach Geschlecht, Studiendauer oder Physikumsnote wurden bei diesen Einschätzungen nicht festgestellt. Diese Kontrollvariablen hatten in derselben Befragung hingegen Einfluss auf die Gesamtzufriedenheit mit der Vorklinik ge-

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.

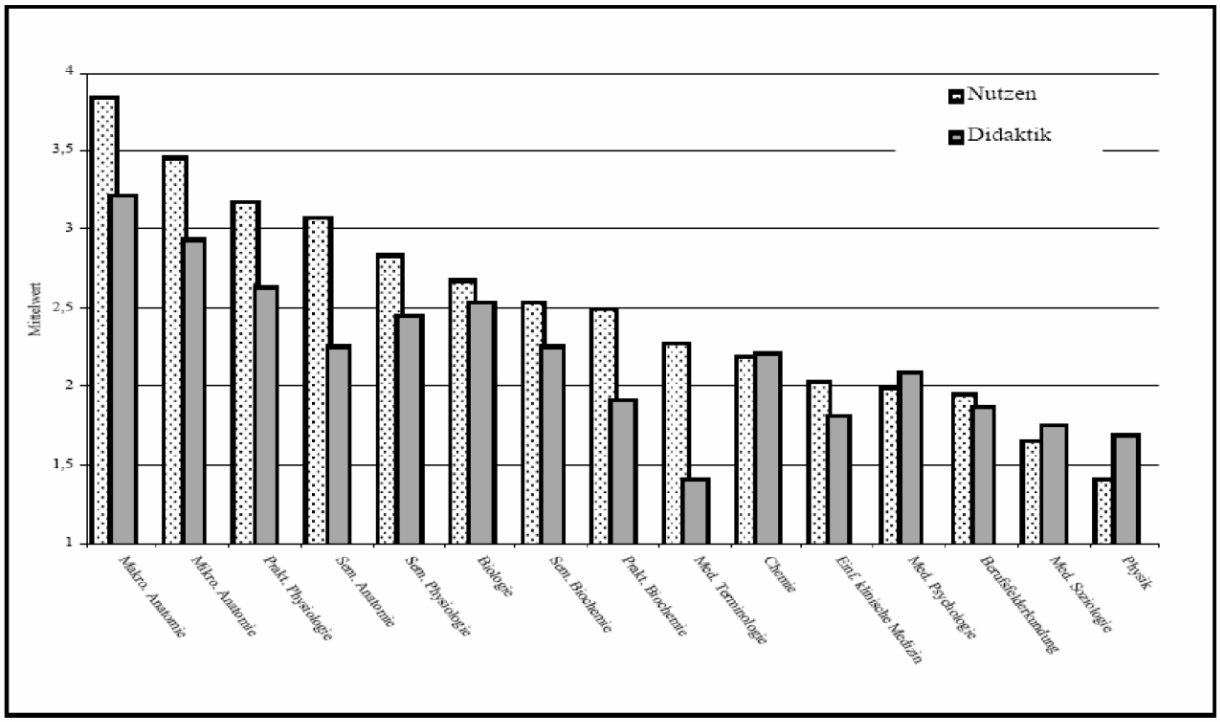

Abbildung 3: Nutzen und didaktische Qualität der vorklinischen Fächer (Mittelwerte, Hamburg 2001)

habt, d.h. Studierende, die durch das Physikum gefallen waren und länger studiert hatten, waren insgesamt unzufriedener.

1.2. Die klinische Ausbildung

Vergleichbare Diskrepanzen zwischen der Beurteilung der einzelnen Veranstaltungen und des Studienabschnitts wurden auch in Bezug auf die klinische Ausbildung festgestellt. Aus Abbildung 4 geht zum Beispiel hervor, dass die didaktische Qualität des Unterrichts am Krankenbett und der Nutzen dieser Veranstaltung in einer retrospektiven Befragung von PJ-Studierenden im Jahr 2001 (N=92; Rücklaufquote 49%) für die Mehrzahl der Fächer als mittelmäßig (= um den Skalenmittelwert von 2.5) beurteilt wurde.

Die Frage nach dem Nutzen der klinischen Ausbildung für das PJ wurde von den gleichen Studierenden in der gleichen Befragung deutlich negativer beantwortet. Die Aussage "Was ich in der klinischen Ausbildung gelernt habe, war für mein PJ von großem Nutzen", wurde nur von einem Viertel der Studierenden bejaht (M=2.1, SD=0.83, vgl. Abbildung 5).

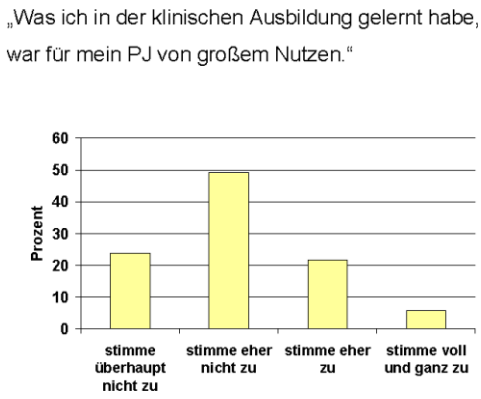

Abbildung 5: Antworthäufigkeiten zum Nutzen der klinischen Ausbildung (M=2.09, SD=0.83)

Als Fazit sei festgehalten, dass die Gesamtbeurteilung eines Studienabschnitts negativer bzw. anders auszufallen scheint als die Beurteilung seiner Bestandteile. Die Erklärung für diese Diskrepanzen liegt vermutlich darin, dass pauschale Fragen eher dazu anregen, auch pauschal zu antworten, während Fragen nach einzelnen Bestandteilen zu einer eher differenzierten Beurteilung führen. Aus Gesamtbeurteilungen von Studienabschnitten oder Studiengängen sollte dementsprechend nur mit Vorsicht auf die einzelnen Fächer bzw. Veranstaltungen geschlossen werden.

• 2. Unterschiede in der Bewertung von Fächern

Implizit wird angenommen, die verschiedenen Fächer hätten vergleichbare Ausgangsbedingungen bei der Bewertung durch Studierende. De facto aber gibt es - und zwar in der Regel unabhängig von der Qualität des Lehrangebots am einzelnen Fakultätsstandort - beliebtere und weniger beliebte Fächer.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.

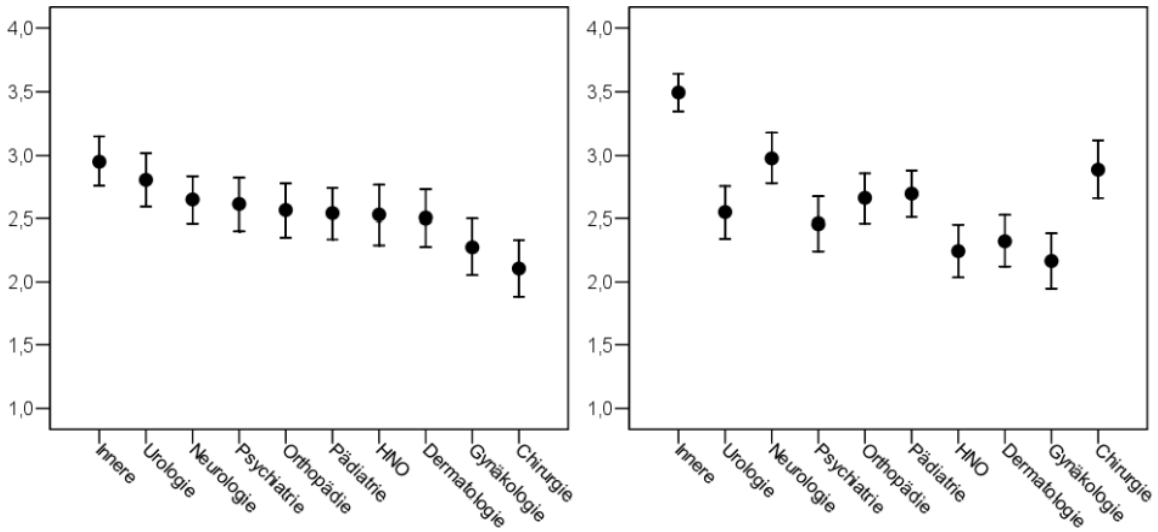

Abbildung 4: Didaktische Qualität (links) und Nutzen (rechts) des Unterrichts am Krankenbett (Mittelwerte und Konfidenzintervalle, Hamburg 2001). (Hier beispielhaft Mittelwertvergleich des Nutzens der einzelnen Fächer mit dem Nutzen der klinischen Ausbildung insgesamt (Abb. 5) mit dem T-Test für gepaarte Stichproben: HNO, Dermatologie und Gynäkologie: $p>0.05$, Urologie: $p=0.003$, Psychiatrie: $p=0.024$, restliche Fächer: $p<0.001$.).

Abbildung 6 zeigt als Beispiel die retrospektive Beurteilung der "didaktischen Qualität" und des "Nutzens für Ihre klinische Ausbildung" der klinisch theoretischen Fächer in der bereits in Abschnitt 1.2. beschriebenen Befragung von PJ Studierenden. Die Abbildung zeigt die beträchtlichen, in der Mehrzahl statistisch signifikanten Unterschiede zwischen den Fächern. Humangenetik, Ökologischer Kurs und Biomathematik liegen in Bezug auf beide Dimensionen unter dem Skalenmittelwert von 2.5, während die anderen Fächer positiv beurteilt wurden.

In den Veranstaltungsgruppen, die in den Abbildungen 4 und 6 dargestellt werden, finden wir in der Regel eine positive Korrelation zwischen der Beurteilung der didaktischen Qualität und der Nutzeneinschätzung der einzelnen Fächer in einer Größenordnung von $r=0.25$ bis $r=0.65$. Diese sind ausnahmslos zweiseitig auf dem 0.05%-Niveau, in den meisten Fällen sogar auf dem 0.01%-Niveau, signifikant. Die logische Interpretation wäre, dass didaktisch gute Veranstaltungen auch einen höheren Nutzen haben. In Einzelfällen unterscheiden sich Nutzen und didaktische Qualität jedoch beträchtlich. Bei einem zweiseitigen t-Test für gepaarte Stichproben auf dem 0.01%-Niveau (Ausschluss listwise, $n=64$) ergeben sich signifikante Unterschiede zwischen den Bewertungen von didaktischer Qualität und Nutzen in folgender Weise:

- Didaktische Qualität signifikant ($p<0.01$) größer als Nutzen: Humangenetik und Ökologischer Kurs.

- Nutzen signifikant ($p<0.01$) größer als didaktische Qualität: Pharmakologie, Pathologie, Innere Medizin, Neurologie und Chirurgie.

Die Auflistung zeigt, dass die Studierenden in vielen Fällen den Nutzen wesentlich höher bewerten als die didaktische Qualität. Offenbar gehen in ihren Nutzenbegriff nicht nur der unmittelbare Nutzen durch die lokale Veranstaltung, sondern auch die lehrangebotsunabhängige Bedeutung des Faches für die Qualifikation als (junger) Arzt ein. Daraus ergibt sich, dass es problematisch ist, nach dem Nutzen eines Faches bzw. einer Veranstaltung zu fragen, wenn man das lokale Lehrangebot beurteilen und bewerten will.

Zusammenhänge zwischen Nutzen und Didaktik müssen auch bezüglich der Bewertung der vorklinischen Fächer in Abbildung 3 diskutiert werden. Der deutliche Zusammenhang zwischen Nutzeneinschätzung und Bewertung der didaktischen Qualität der Fächer könnte auch in diesem Fall darauf zurückzuführen sein, dass gegenständliche und körperbezogene Fächer grundsätzlich bevorzugt werden, was die positive Einschätzung der Fächer Anatomie, Physiologie und Biologie (mit-)bedingen würde. Womöglich hängt dies auch damit zusammen, dass diese Fächer von ihrer "Natur" her häufiger und plastischer in der Lage sind, dem studentischen Desideratum nach klinischen Bezügen im Unterricht zu entsprechen. Diese These wird durch mehrere Evaluationen der vorklinischen Ausbildung gestützt, in denen das Fach Anatomie durchweg gute Beurteilungen erhielt während Chemie, Physik, Berufsfelderkundung und Medizinische Psychologie eher negativ beurteilt wurden [10][11][12]. Es stellt sich aber auch die Frage nach der Fähigkeit der Studierenden, zwischen Nutzen und Didaktik zu differenzieren. Die unterschiedliche Beurteilung von didaktischer Qualität und Nutzen einzelner Fächer lässt sich als Beleg für die differenzierte Urteilsfähigkeit interpretieren, die über alle Fächer hinweg deutliche Korrelation hingegen als Beleg dagegen.

Das Problem eines möglichen Fächerbias stellt sich mit besonderer Schärfe, wenn aus den studentischen Bewertungen der Fächer im Rahmen von Konzepten zur leistungsorientierten Mittelvergabe Fächerrankings und darauf basierend Mittelallokationen vorgenommen werden. In diesem Fall müssten für die nachweislich attraktiveren Fächer "Handicaps" eingebaut werden.

Für die Fragebogenkonstruktion heißt dies, dass in den Erhebungen nur nach Parametern gefragt werden sollte, die vom lokalen Lehrangebot und nicht von allgemeinen studentischen Präferenzen bzw. Vorstellungen über den Arztberuf abhängen. Dies bedeutet konkret, dass vorrangig nach Parametern der Prozessqualität des lokalen Unterrichts, wie Dozentenverhalten, Organisationsqualität, Brauchbarkeit von Lernmaterialien etc. gefragt werden sollte und nicht nach möglicherweise nicht lokal bedingten Faktoren wie Nutzen für den ärztlichen Beruf.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.

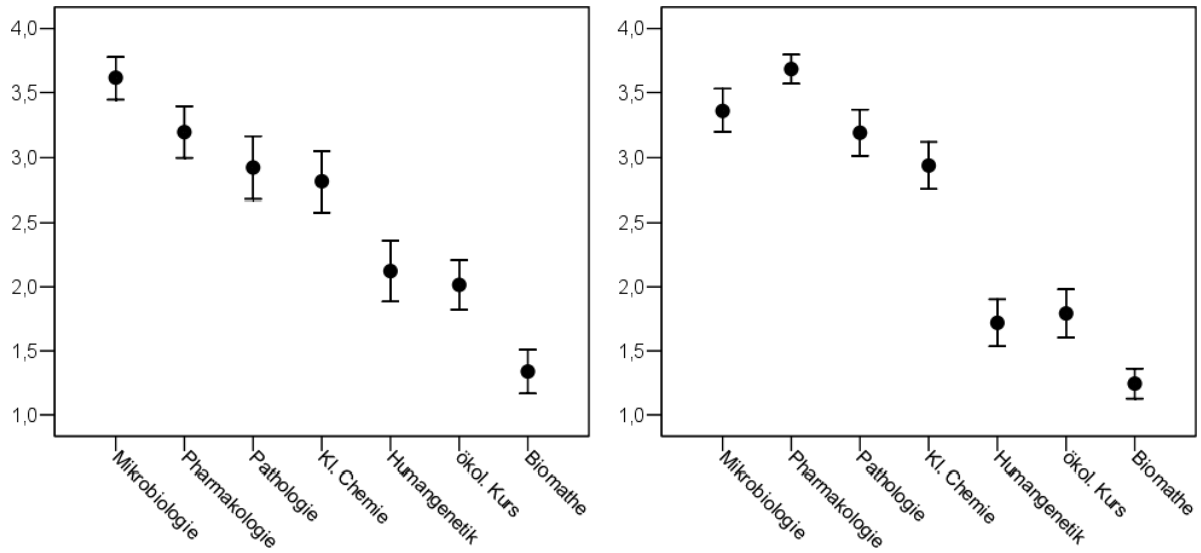

Abbildung 6: Didaktische Qualität (links) und Nutzen (rechts) der klinisch-theoretischen Fächer (Mittelwerte und Konfidenzintervalle, Hamburg 2001). Alle Unterschiede zwischen den Fächern sind, mit Ausnahme derjenigen zwischen Ökologischem Kurs und Humangenetik bzw. zwischen Pathologie und Klinischer Chemie, statistisch signifikant; Chi²-Berechnungen auf Basis der Absolutzahlen der extremen Skalenwerte 1 und 4.

Allerdings sind bezüglich der Attraktivitätsunterschiede auch andere Zusammenhänge denkbar:

• Es könnte postuliert werden, dass Fächer, bei denen der Nutzen weit größer als die didaktische Qualität eingeschätzt wird, tatsächlich auch diejenigen sind, bei denen didaktische Verbesserungen am ehesten angezeigt wären. Umgekehrt finden sich auch immer wieder prinzipiell weniger attraktive Fächer, die sich durch besondere didaktische Anstrengungen gute Beurteilungen sichern.

• Unterschiede in der Beurteilung des Nutzens bzw. der didaktischen Qualität von Fächern könnten auch mit der relativen Angebotsmenge im Sinne einer Mitbeurteilung der Darbietungsdauer ("Quantität anstatt Qualität") zusammenhängen. So korrelieren in der beschriebenen Vorklinik-Untersuchung die Zahl der Veranstaltungsstunden pro Fach mit den Mittelwerten der Beurteilung des Nutzens mit r=0.39 bzw. mit der didaktischen Qualität mit r=0.59. Damit wäre die Beurteilung des Nutzens und der Qualität der Veranstaltungen (auch) ein Spiegelbild der Schwerpunkte des jeweiligen Studienangebots.

• 3. Beliebte und ungeliebte Veranstaltungsformen

An der Kohorte der Hamburger Vorklinik Absolventen (Medizinstudierende im 5. klinischen Semester) wurde 2001 auch überprüft, welche Studierenden die jeweiligen Veranstaltungsformen - Vorlesung, Seminar oder Praktikum - bevorzugten, wovon diese Präferenzen abhängen und in welchem Zusammenhang sie mit dem subjektiv empfundenen Lernerfolg stehen. Die Auswertung zeigt, dass Praktika und Seminare gegenüber Vorlesungen im Durchschnitt als deutlich nützlicher angesehen wurden. Diese unterschiedliche Bewertung der Veranstaltungsformen findet sich auch in den Antworten auf die Frage wieder, ob der Stundenumfang einer Veranstaltungsform verringert, beibehalten oder vergrößert werden sollte. So wünschten sich weit mehr Studierende einen Ausbau der Seminare (26%) und Praktika (25%), als dies für Vorlesungen (2%) der Fall war. In einer vergleichbaren Erhebung im Jahr 2004 wurde die negative Beurteilung der Vorlesung bestätigt: 39% der Studierenden wünschten einen Ausbau der Seminare, 22% einen Ausbau der Praktika, aber nur 12% einen solchen der Vorlesungen.

Hieraus folgt, dass Fächer, die relativ stärker auf Vorlesungen basieren, in der studentischen Bewertung in der Regel von vornherein negativer bewertet werden dürften als solche, die andere Veranstaltungsformen, insbesondere kommunikativere oder praxisorientiertere Typen, anbieten. Hieraus folgt auch, dass es bei vergleichenden, insbesondere rankenden Evaluationen sinnvoll erscheint, vergleichbare Veranstaltungsformen - z.B. Vorlesungen mit Vorlesungen, Unterricht am Krankenbett mit Unterricht am Krankenbett miteinander zu vergleichen bzw. bei Fächervergleichen auf den möglichen Bias der Veranstaltungsform zu achten.

• 4. Mögliche Unterschiede zwischen zeitnahen und zeitfernen Veranstaltungsevaluationen

Mehrfach ist aufgefallen, dass die Bewertung der gleichen Veranstaltung durch Studierende je nach Bewertungszeitpunkt - zeitnah oder zeitfern - sehr unterschiedlich ausfallen kann. Als Beispiel werden die Evaluationsergebnisse der ersten Durchführung des Querschnittsbereichs "Gesundheitsökonomie, Gesundheitssystem, Öffentliche Gesundheitspflege" nach der neuen Approbationsordnung für Ärzte in Hamburg im Frühjahr 2004 dargestellt.

Das Lehrangebot für diesen Querschnittsbereich umfasste:

• Vier einstündige Einführungsvorlesungen und

• je eine Veranstaltungsreihe zu den Themen "neue Versorgungsformen" einerseits und "Qualitätssicherung" andererseits, die jeweils aus einer Exkursion zu einem Beispiel für eine neue Versorgungsform bzw. einem Qualitätssicherungsprojekt sowie einem Vor- und einem Nachbereitungsseminar bestand.

• Über die Exkursionen hatten die Studierenden in Sechsergruppen einen kritischen Bericht und eine mündliche Präsentation am Overheadprojektor anzufertigen.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.

Da auch die Eignung der einzelnen Exkursionsprojekte erfasst werden sollte, wurde zeitnah, d.h. im Anschluss an die beiden Nachbereitungsseminare, eine Evaluation durchgeführt. Darüber hinaus bewerteten die gleichen Studierenden alle Veranstaltungen am Ende des jeweiligen Trimesters, d.h. nach ca. 12 Wochen Unterrichtsbetrieb. Letztere Evaluation umfasste alle Fächer und Veranstaltungen des Blockes, in casu fünf Fächer und zwei Querschnittsbereiche.

Tabelle 1 fasst die Ergebnisse der Evaluationen zusammen. In den beiden zeitnahen Evaluationen (N=151 Fragebögen bei 85 Teilnehmern - jeder Studierende hätte idealerweise zwei Evaluationsbögen abgeben sollen, einen für Neue Versorgungsformen und einen für Qualitätssicherung) schnitten die Seminare des Querschnittsbereiches gut, die Exkursionen mittelmäßig ab. In der Endevaluation (n=72) ergab sich aber überraschenderweise ein insgesamt deutlich negativeres Bild. Zufriedenheit und Lernerfolg bezüglich der Seminare wurden kapp negativ beurteilt, die Zufriedenheit mit den Exkursionen, Hausarbeiten und Präsentationen wurde nunmehr als sehr gering angegeben (M=2.3). Im Großen und Ganzen ergab sich eine Differenz zwischen zeitnaher und zeitferner Evaluation von einem ganzen Punkt auf der Sechserskala.

Tabelle 1: Beurteilung des Querschnittsbereiches Gesundheitsökonomie, Mittelwerte (SD) der Angaben auf einer Sechserskala (1=niedrig, 6=hoch)

Item	zeitnahe Evaluation	zeitferne Evaluation
Qualität Vorbereitung Dozenten		4.2 (1.5)
Didaktische Qualität Seminare	4.2 (1.2)	3.6 (1.3)
Subjektiver Lernerfolg Seminare	3.8 (1.2)	3.0 (1.2)
Zufriedenheit Seminare	Vorbereitungsseminare 4.0 (1.1) Nachbereitungsseminare 4.2 (1.1)	3.0 (1.4)
Subjektiver Lernerfolg Exkursionen	3.5 (1.3)	
Weiterempfehlung Exkursionen	3.4 (1.4)	
Zufriedenheit Exkursionen, Hausarbeiten und Präsentationen		2.3 (1.4)

Viele Erklärungen für diese Unterschiede zwischen zeitferner und zeitnaher Evaluation sind denkbar. Zu der positiven Bewertung, die direkt im Anschluss an die Nachbereitungsseminare abgegeben wurde, mag die Erleichterung der Studierenden beigetragen haben, einen guten Vortrag gehalten zu haben. Denn 79 bzw. 88 Prozent der Präsentationen in den Unterrichtseinheiten Neue Versorgungsformen bzw. Qualitätssicherung wurden mit der Note 1 oder 2 bewertet und den Studierenden sofort mitgeteilt. Auf der anderen Seite kann die eher negative Bewertung der Veranstaltungen in der zeitfernen Evaluation folgende Gründe gehabt haben:

- Im Vergleich zu anderen, noch positiver erlebten Veranstaltungen des Themenblocks, könnten die Studierenden die Bewertung des Querschnittsbereiches relativiert haben.

- Die negative Bewertung könnte das Resultat einer insgesamt schlechten Stimmung der Studierenden am Ende des Trimesters sein. Sie wäre somit Ausdruck einer pauschalen Unzufriedenheit, die sich - anders als in Abschnitt 2.1 beschrieben - auch in der Bewertung der einzelnen Veranstaltungen niederschlägt.

- Sie könnte aber auch die Verärgerung über die vielen Leistungsnachweise im Querschnittsbereich widerspiegeln. So hatte jeder Studierende 2 Präsentationen und 2 Hausarbeiten anfertigen müssen und bereitete sich gegen Ende des Trimesters - zum Zeitpunkt der zeitfernen Evaluation - auf die abschließende Klausur vor.

Die Ursache für die Diskrepanz zwischen zeitnaher und zeitferner Evaluation ist im Nachhinein nicht zu ermitteln. Es empfiehlt sich allerdings bei der Evaluation von Veranstaltungen unterschiedliche Zeitpunkte zu wählen und die Kontextabhängigkeit des Bewertungsverhaltens zu berücksichtigen.

• 5. Beurteilungsunterschiede und Studiendauer

Ein vergleichbares Problem des Befragungszeitpunktes dürfte es auch für das Gesamtstudium geben. Zur Beurteilung der Qualität des Studiengangs, d.h. losgelöst von einzelnen Lehrveranstaltungen und Dozenten, untersuchten Bargel und Ramm [13] die Antworten auf einige globale Items zu inhaltlicher (fachliche Güte des Lehrangebots), struktureller (Aufbau und Gliederung), didaktischer (Art und Weise der Veranstaltungsdurchführungen) und tutoraler (Betreuung und Beratung) Qualität des Studiengangs Medizin. Im Vergleich der Fachsemestergruppen untereinander schnitt das Medizinstudium bei den Studierenden der Vorklinik am besten ab. Je länger die Studierenden studiert hatten, desto schlechter fiel ihre Beurteilung der Erfahrungen im bisherigen Studium aus (vgl. Tabelle 2). In der wissenschaftlichen Literatur ist die Frage, ob Veranstaltungen, die später im Studium stattfinden, grundsätzlich schlechter oder besser bewertet werden, allerdings nicht abschließend geklärt. Marsh und Roche [4] sowie D'Appolonia und Abrami [5] resümieren in ihren Übersichtsarbeiten, dass die diesbezüglichen Studienergebnisse uneinheitlich und die Effekte gering sind.

Tabelle 2: Bewertung der Studienqualität im Fach Humanmedizin nach Fachsemestern (WS 1992/93, [13])

	Bewertung der Grundelemente der Studienqualität			
	Inhalt	Aufbau	Didaktik	Betreuung
1.-4. Fachsemester	0.7	-0.2	0.0	-0.8
5.-8. Fachsemester	0.3	-0.2	-0.4	-1.1
9.-12. Fachsemester	-0.2	-0.6	-1.1	-1.6
Insgesamt (inkl. ≥ 13. Fachsemester)	0.2	-0.5	-0.6	-1.3

Dargestellt sind Mittelwerte, Antwortskala rangierte von –3 "sehr schlecht" bis +3 "sehr gut"

Diskussion und Schlussfolgerungen

Die Analyse hat gezeigt, dass die Ergebnisse standardisierter Evaluationen der ärztlichen Ausbildung eine Reihe von Erhebungs- und Interpretationsproblemen in sich bergen. Dementsprechend ist stets vorsichtig beim Ziehen von Schlussfolgerungen bzw. gar beim Ranking von Fächern und Veranstaltungen zu verfahren. Eine Verbesserung der Interpretationssicherheit kann über folgende Gesichtspunkte erreicht werden:

- Aus Gesamtbeurteilungen sollte nur mit Vorsicht auf die einzelnen Veranstaltungen geschlossen werden.

- Bei Rankings sind "Handicaps" der theoretischen bzw. patientenfernen Fächer zu bedenken bzw. im Sinne einer besseren Vergleichbarkeit „Handicaps" für klinische Fächer einzubauen.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.

- Einmalige Messungen reichen für valide Aussagen nicht aus, Mehrfachmessungen, auch Längsschnittvergleiche, sind eine Grundvoraussetzung.

- Vergleiche sollten nach Möglichkeit nur bei vergleichbaren Einheiten (z.B. nach Veranstaltungstyp) vorgenommen werden.

- Empirisch gefundene Unterschiede sind immer auf ihre Ursachen(-Bündel) zu untersuchen. Auch Fakultätsvergleiche können helfen, Fehlinterpretationen zu vermeiden.

- Gleiches gilt für die Notwendigkeit, die Bewertungen nicht von einer Person, sondern aus verschiedenen Blickwinkeln (z.B. von gemischten Evaluationskommissionen) vorzunehmen.

- Veranstaltungen und Fächer sollten grundsätzlich nach verschiedenen Gesichtspunkten bewertet werden. Für lokale Entscheidungsprozesse sollte nur nach lokal beeinflussbaren Faktoren gefragt werden: Erhebungen zur Prozessqualität sind hier grundsätzlich brauchbarer als solche zur Ergebnisqualität.

Zusammenfassend betrachtet gelten für die Evaluation von Ausbildungsangeboten die Grundannahmen der Triangulation [14] beim Einsatz von Forschungsinstrumenten. Mehrere Zugangswege - qualitative und quantitative Methoden, offene und geschlossene Erhebungstechniken - sind zu wählen und die auf verschiedenen methodischen Wegen gewonnenen Aussagen zu gleichen Fragestellungen vergleichend zu analysieren, um auf diese Weise ein Maximum an Aussagekraft zu erhalten.

Zuallerletzt: Die beschriebenen Probleme sind kein Grund, Evaluationen der Ausbildung zu unterlassen. Perfektion gibt es hier genauso wenig bei der Beurteilung von Forschungsleistungen oder bei der Evaluation von Versorgungsprozessen. Die weitere Beforschung von Ansätzen zur Lösung der beschriebenen Evaluationsprobleme ist dringend notwendig.

Korrespondenzadresse:

• Dr . med. Hanna Kaduszkiewicz, Universitätsklinikum Hamburg-Eppendorf, Institut für Allgemeinmedizin, Zentrum für Psychosoziale Medizin, Martinistraße 52, 20246 Hamburg, Deutschland, Tel. 040/42803-3247, Fax.: 040/42803-3681 kaduszki@uke.uni-hamburg.de

Literatur:

[1] Pabst R. Lehrevaluation in der Medizin. Befragungen zur Qualität der Lehre in den Hochschulen. Dtsch Arztebl. 2001; B630-B632.

[2] Theisel N, Stosch C, Koebke J. Evaluationsbemühungen an den Medizinischen Fakultäten in Deutschland - Ergebnisse einer Umfrage. Med Ausbild. 2000;17:18-21.

[3] Weber A, Wacker A, Weltle D, Lehnert G. Stellenwert der Lehre an den deutschen medizinischen Fakultäten. Dtsch Med Wochenschr. 2000;125:1560-1564.

[4] Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective. Am Psychol. 1997;52:1187-1197.

[5] D'Appolonia S, Abrami PC. Navigating Student Ratings of Instruction. Am Psychol. 1997;52:1198-1208.

[6] McKeachie WJ. Student Ratings. The Validity of Use. Am Psychol. 1997;52:1218-1225.

[7] Marsh HW. Student's Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. Br J Educ Psychol. 1984;76:707-754.

[8] Bortz J, Döring N. Forschungsmethoden und Evaluation. Berlin. Heidelberg. New York: Springer-Verlag. 1995.

[9] Kohler N, van den Bussche H. Je schwieriger desto beliebter. Nutzen, didaktische Qualität und Schwierigkeitsgrad des vorklinischen Lehrangebots aus der Sicht von Hamburger Medizinstudenten. Ann Anat. 2004;186:283-288.

[10] Pabst R, Rothkötter HJ. Retrospective evaluation of a medical curriculum by final-year students. Med Teach. 1996;18:288-293.

[11] Pabst R, Rothkötter HJ. Retrospective Evaluation of Undergraduate Medical Education by Doctors at the End of their Residency Time in Hospitals: Consequences for the Anatomical Curriculum. Anat Rec. 1997;249:431-434.

[12] Medizinische Fakultät der Universität zu Köln. Lehrbericht der Medizinischen Fakultät der Universität zu Köln. Köln: Universität Köln. 1999.

[13] Bargel T, Ramm M. Das Studium der Medizin. Eine Fachmonographie aus studentischer Sicht. Schriftenreihe Studien zu Bildung und Wissenschaft 118. Bonn: Bundesministerium für Bildung und Wissenschaft. 1994.

[14] Dunkelberg H, van den Bussche H. Triangulation: Von unterschiedlichen Ergebnissen beim Einsatz unterschiedlicher Methoden. Z Arztl Fortbild Qualitatssich. 2004;98:519-525.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000256.shtml
Bitte zitieren als: 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. GMS Z Med Ausbild. 2006;23(2):Doc37.