Implementierung und Erprobung eines Lernziel-basierten Evaluationssystems im Studium der Humanmedizin

Tobias Raupach[1,2]
Sarah Schiekirka[3]
Christian Münscher[3]
Tim Beißbarth[4]
Wolfgang Himmel[5]
Gerhard Burckhardt[6]
Tobias Pukrop[7]

1 Universitätsmedizin Göttingen, Abteilung Kardiologie & Pneumologie, Göttingen, Deutschland

2 University College London, Department of Clinical, Educational and Health Psychology, London WC1E 7HB, United Kingdom

3 Universitätsmedizin Göttingen, Studiendekanat, Göttingen, Deutschland

4 Universitätsmedizin Göttingen, Abteilung Medizinische Statistik, Göttingen, Deutschland

5 Universitätsmedizin Göttingen, Abteilung Allgemeinmedizin, Göttingen, Deutschland

6 Medizinische Fakultät Göttingen, Studiendekan, Göttingen, Deutschland

7 Universitätsmedizin Göttingen, Abteilung Hämatologie & Onkologie, Göttingen, Deutschland

Zusammenfassung

Zielsetzung: Aktuell werden an den deutschen medizinischen Fakultäten unterschiedliche Konzepte zur leistungsorientierten Mittelvergabe (LOM) in der Lehre diskutiert. Die Umsetzung scheitert mitunter am Mangel valider Messkriterien zur Beurteilung der Lehrqualität. Neben der Struktur und den Prozessen der Lehre sollte das Ergebnis der Lehre im Mittelpunkt der Qualitätsbewertung stehen. Ziele dieser Arbeit waren die Erprobung eines neuen, lernzielbezogenen Evaluationssystems im klinischen Abschnitt des Studiums der Humanmedizin und der Vergleich der Ergebnisse mit den Daten eines traditionellen Evaluationsverfahrens.
Methodik: Aus studentischen Selbsteinschätzungen zu Beginn und Ende eines jeden Lehrmoduls wurde nach einer neu entwickelten Formel der lernzielbezogene, prozentuale Lernerfolg berechnet. Die Lernerfolgs-Mittelwerte pro Modul wurden mit traditionellen Evaluationsparametern, insbesondere mit Globalbewertungen, ins Verhältnis gesetzt.
Ergebnisse: Der mittels vergleichender Selbsteinschätzungen berechnete Lernerfolg und die Globalbewertungen produzierten deutlich unterschiedliche Rangfolgen der 21 klinischen Module. Zwischen dem Lernerfolg und den Globalbewertungen fand sich keine statistisch signifikante Korrelation. Allerdings korrelierten die Globalbewertungen stark mit den studentischen Erwartungen vor Modulbeginn und mit strukturellen und prozeduralen Parametern der Lehre (Pearson's r zwischen 0,7 und 0,9).
Schlussfolgerung: Die Messung des Lernzuwachses mittels vergleichender studentischer Selbsteinschätzungen kann die traditionelle Evaluation um eine wichtige Dimension erweitern. Im Unterschied zu studentischen Globalbewertungen ist das neue Instrument lernzielbezogen und unabhängiger vom Einfluss Konstrukt-irrelevanter Parameter. Hinsichtlich der Entwicklung eines LOM-Algorithmus eignet sich das neue Instrument gut zur Beurteilung der Lehrqualität.

Schlüsselwörter: Evaluation, Selbsteinschätzung, LOM, Lernziel, Kongruenz, klinisches Studium

Einleitung

Auf der Grundlage der Anregungen des Wissenschaftsrates [1] fand im Jahr 2009 an über der Hälfte aller Medizinischen Fakultäten in Deutschland eine leistungsorientierte Mittelvergabe (LOM) für gute Lehre statt.
Auch die Verteilung von Landesmitteln auf einzelne Fakultäten orientiert sich zunehmend an Leistungsparametern. Diesbezüglich wurden in Nordrhein-Westfalen verschiedene Qualitätskriterien erarbeitet, die in drei Ebenen zusammengefasst wurden [2]: Struktur-, Prozess- und Ergebnisqualität der Lehre. Während die Evaluation der Strukturen (Ressourcen, Stundenpläne) und Prozesse (didaktische Gestaltung der Lehre, Prüfungen) nicht im Mittelpunkt der Betrachtungen stand, schlugen die Autoren vor, zur Beurteilung des Studienergebnisses globale Parameter wie die studentischen Leistungen im Staats-

examen, die Studiendauer und die Retentionsrate heranzuziehen. Diese Kriterien sind insbesondere im klinischen Studienabschnitt nicht nach Modulen und Fächern aufgeschlüsselt und eignen sich somit praktisch nicht zur Verteilung von LOM-Mitteln innerhalb einer Fakultät. Es ist somit notwendig, Qualitätsindikatoren zu entwickeln, die zum Ranking von Lehrveranstaltungen oder Fächern innerhalb einer Fakultät herangezogen werden können. Diese Indikatoren sollten sich auch auf Fakultätsebene auf Strukturen, Prozesse und das Ergebnis der Lehre beziehen.

Eine Erhebung im Jahre 2009 ergab, dass an vielen Fakultäten studentische Evaluationen die Grundlage für die Zuteilung von LOM-Lehre waren [3]. Gebräuchliche Evaluationsbögen beinhalten überwiegend Fragen zu strukturellen und organisatorischen Aspekten sowie globale Einschätzungen auf Schulnotenskalen. Zwar erlauben diese Instrumente möglicherweise eine Beurteilung von Strukturen und Prozessen; inwieweit hieran aber auch die Ergebnisqualität abgelesen werden kann, ist bislang unklar.

Wenngleich verschiedene Definitionen des gewünschten „Ergebnisses“ im Medizinstudium denkbar sind, liegt es nahe, den studentischen Lernerfolg als eine mögliche Messgröße für das Studien-Ergebnis festzulegen. Wünschenswert wäre folglich ein Evaluationsinstrument, mit dem der in einer Veranstaltung erzielte Lernerfolg abgeschätzt werden kann. An der Medizinischen Fakultät Göttingen wurde kürzlich ein entsprechendes lernzielbasiertes Evaluationsinstrument entwickelt, dessen Reproduzierbarkeit und *Kriteriums*-Validität belegt werden konnten [4]. Somit steht neben der traditionellen Evaluation mit den Hauptzielgrößen „Struktur- und Prozessqualität“ nun auch ein neues Instrument mit der Zielgröße „Lernerfolg“ zur Verfügung. Zur Untersuchung der *diskriminanten* Validität des Instruments sollten in der vorliegenden Studie folgende Forschungsfragen beantwortet werden:

1. Ergibt sich auf der Grundlage des lernzielbasierten Instruments eine andere Rangfolge der Lehrveranstaltungen als auf der Grundlage traditioneller Evaluationsparameter? Es wird die Hypothese aufgestellt, dass durch beide Verfahren unterschiedliche Aspekte der Lehre erfasst werden und sich daher eine Rangfolge der Lehrveranstaltungen nach strukturellen/prozeduralen Kriterien deutlich von einer Rangfolge der Veranstaltungen nach dem hierin erzielten Lernerfolg unterscheidet.
2. Inwieweit korrelieren die Erwartungshaltung der Studierenden und traditionelle Evaluationsparameter untereinander sowie mit den Ergebnissen der lernzielbezogenen Evaluation? Es wird die Hypothese formuliert, dass traditionelle Evaluationsparameter untereinander starke Korrelationen aufweisen. Da Strukturen und Prozesse ebenfalls einen Einfluss auf die Lehrqualität ausüben und somit auch den Lernerfolg beeinflussen können, werden zwischen den Ergebnissen der beiden Evaluationsinstrumente moderate Korrelationen erwartet.

Methodik

Klinisches Curriculum und Evaluationsinstrumente an der Medizinischen Fakultät Göttingen

Der dreijährige klinische Studienabschnitt in Göttingen ist modular organisiert. Die 21 interdisziplinären Module erstrecken sich über zwei bis sieben Wochen. Im ersten klinischen Jahr werden ärztliche Basisfertigkeiten sowie die Grundlagen der Infektiologie und Pharmakologie vermittelt, in den folgenden drei Semestern werden Gesundheitsstörungen samt ihrer Behandlung thematisiert und im letzten klinischen Semester differentialdiagnostische Aspekte betont.

1. Traditionelle Evaluation: Jeweils zu Modulende findet eine Online-Evaluation statt (EvaSys®, Electric Paper, Lüneburg). Hierbei bewerten die Studierenden organisatorische und strukturelle Aspekte der Lehre auf einer sechsstufigen Skala. Die vom Evaluationsausschuss der Fakultät entwickelten Aussagen lauten:
   1.1. „Die Umsetzung des interdisziplinären Unterrichts ist in diesem Modul sehr gut gelungen.“ (Interdisziplinarität)
   1.2. „In diesem Modul wurde das selbständige Aufarbeiten der Lernziele gefördert.“ (Selbstlernkompetenz)
   1.3. „Bezogen auf meine berufliche Zukunft schätze ich meinen Lernzuwachs in diesem Modul als sehr hoch ein.“ (subjektiver Lernzuwachs)
   1.4. „Mit der grundlegenden Struktur des Moduls (Aufbau, Unterrichtsformen, Zeitplan) war ich sehr zufrieden.“ (Modulstruktur)
   1.5. „Dieses Modul sollte so fortgeführt werden wie bisher.“ (Beibehaltung)
2. Schließlich werden die Studierenden mit dem Item „Bitte bewerten Sie das Modul mit einer Gesamtnote“ um eine Globalbewertung nach dem Schulnotensystem gebeten.
3. Um einen möglichen Zusammenhang zwischen studentischen Erwartungen vor Modulbeginn und Bewertungen nach Modulende zu untersuchen, wurde zum Zweck dieser Studie für jedes Modul eine Online-Vorbefragung eingerichtet. Hierbei sollten die Studierenden auf einer sechsstufigen Skala angeben, inwieweit sie den folgenden drei Aussagen zustimmen:
   3.1. „Ich glaube, dieses Modul ist wichtig für meine spätere berufliche Zukunft.“ (Wichtigkeit)
   3.2. „Ich freue mich schon auf dieses Modul.“ (Vorfreude)
   3.3. „Ich habe von Kommilitoninnen und Kommilitonen vor allem Gutes über dieses Modul gehört.“ (Ruf des Moduls)

4. Lernzuwachs-Evaluation mittels vergleichender Selbsteinschätzung (VSE): Das neu entwickelte Evaluations-Instrument bestimmt auf der Grundlage wiederholter studentischer Selbsteinschätzungen den Lernzuwachs für spezifische Lernziele. In der Online-Befragung vor Modulbeginn wurden die Studierenden gebeten, ihr Leistungsniveau bezüglich spezifischer Lernziele (Beispiel: „Ich kann ein EKG interpretieren.“) auf einer sechsstufigen Skala von „trifft voll zu“ bis „trifft überhaupt nicht zu“ einzuschätzen. Diese Selbsteinschätzung wurde in der unverändert durchgeführten Online-Evaluation zu Modulende wiederholt. Mittels Division der Mittelwert-Differenz (prä-post) für ein spezifisches Lernziel durch den korrigierten Mittelwert der initialen Selbsteinschätzungen aller Studierenden wurde der lernzielbezogene Lernzuwachs als Prozentwert berechnet (siehe Abbildung 1):

$$\text{Lernzuwachs [\%]} = \frac{\mu_{prä} - \mu_{post}}{\mu_{prä} - 1} \times 100$$

($\mu_{prä}$ = durchschnittliche initiale Selbsteinschätzung; μ_{post} = durchschnittliche Selbsteinschätzung nach dem Modul)

Abbildung 1: Lernzuwachs [%]

Der Gesamt-Lernzuwachs eines Moduls wurde als Mittelwert aus den prozentualen Lernzuwachs-Werten von 15 spezifischen Lernzielen berechnet, die für jedes Modul entsprechend den Vorgaben des Göttinger Lernzielkatalogs (http://www.med.uni-goettingen.de/de/media/G1-2_lehre/lernzielkatalog.pdf) formuliert wurden. Zwischen dem oben aufgeführten Evaluations-Item „Bezogen auf meine berufliche Zukunft schätze ich meinen Lernzuwachs in diesem Modul als sehr hoch ein“ (subjektiver Lernzuwachs) und dem Lernzuwachs auf dem Boden der VSE bestehen wesentliche Unterschiede: Der subjektive Lernzuwachs wird zu einem einzigen Zeitpunkt erhoben und hat globalen Charakter. Der Lernzuwachs nach VSE wird aus zwei Messpunkten errechnet und bezieht sich auf spezifische Inhalte. In einer kürzlich publizierten Reliabilitäts- und Validitätsstudie konnte gezeigt werden, dass der so berechnete Lernzuwachs sehr gut mit objektiven Leistungsparametern korreliert [4].

Stichprobenbeschreibung

Im Wintersemester 2008/09 waren insgesamt 977 Studierende zur Teilnahme an den Modulen des klinischen Studienabschnitts angemeldet. Alle Studierenden wurden per E-Mail zur Teilnahme an der Evaluation eingeladen; über ein automatisches Versandsystem erhielt jeder Studierende pro Befragung drei E-Mails mit einem direkten und nur einmal verwendbaren Link zur Online-Plattform. Für die Befragungen zu Modulbeginn wurde die Datensammlung drei Tage vor Modulbeginn gestartet und drei Tage nach Modulbeginn beendet; analoge Zeiträume wurden für die Befragungen zu Modulende vorgegeben. Die Teilnahme an der Evaluation war freiwillig, und alle Daten wurden anonym eingegeben. Daher ist keine nähere Charakterisierung der Stichprobe nach Alter und Geschlecht möglich.

Datenerhebung und -analyse

In die vorliegende Arbeit gingen anonym gesammelte Evaluationsdaten ein, die im Wintersemester 2008/09 erhoben wurden. Die Studierenden wurden gebeten, in der Modul-Abschlussevaluation anzugeben, ob sie auch an der Eingangsbefragung teilgenommen hatten. Um Prä-Post-Vergleiche zu ermöglichen, wurden lediglich Daten von Studierenden verwendet, die nach eigenen Angaben an beiden Befragungen teilgenommen hatten. Da die Datenanalyse sich auf Mittelwerte ganzer Studierendenkohorten bezieht, war keine individuelle Kennung von Studierenden erforderlich.

Zur Beantwortung der ersten Forschungsfrage wurden Ranglisten der 21 Module erstellt, die sich entweder auf den Mittelwert der Globalbewertung oder den über 15 Lernziele gemittelten Lernzuwachs (VSE) bezogen. Zur Bearbeitung der zweiten Fragestellung wurden die Korrelationen zwischen den Mittelwerten der traditionellen Evaluationsparameter und dem mittleren Lernzuwachs (VSE) in den Modulen untersucht. Zur statistischen Datenanalyse wurde SPSS® 14.0 (Illinois, USA) verwendet. Alle Daten zeigten sich im Kolmogorov-Smirnov-Anpassungstest normalverteilt. Folglich sind Korrelationen als Pearson's r angegeben. Der quadrierte Korrelationskoeffizient gibt dabei das Bestimmtheitsmaß der Korrelation und somit auch die Varianzaufklärung an.

Ergebnisse

Rücklaufquoten

Von den 977 zu den klinischen Modulen angemeldeten Studierenden gaben 573 Studierende insgesamt 51.915 Einzelratings ab. Die Rücklaufquoten innerhalb der verschiedenen Module lagen zwischen 36,7 und 75,4%.

Modulrankings im Vergleich

Tabelle 1 stellt die mittels der traditionellen Evaluation erhobenen Globalbewertungen den Lernzuwachsdaten der verschiedenen Module gegenüber. Die Module wurden gemäß dieser Ergebnisse in zwei Rangreihen gebracht (siehe die letzten beiden Spalten der Tabelle). Teilweise führten die beiden Methoden zu deutlich divergierenden Rangplätzen. Bei sechs der 21 Module wiesen die Ergebnisse Unterschiede von mindestens sechs Rangplätzen auf. Besonders fiel das Modul 19 auf, das mit 1,56 die beste Note erreichte, nach der Lernzuwachs-Evaluation mit durchschnittlich 55,6% jedoch den 17. Platz belegte. Umgekehrt belegte das Modul 7 (Thema „Evidenzbasierte Medizin“) im Ranking nach der Globalbewertung mit einer Gesamtnote von 3,67 den letzten Platz, obwohl in diesem Modul ein mittlerer Lernzuwachs

Tabelle 1: Globalbewertungen und Lernzuwachsdaten sowie die Rangplätze der 21 Module. Die Module sind in der zeitlichen Reihenfolge, wie sie in dem dreijährigen klinischen Curriculum stattfinden, aufgeführt.

Modulnummer (Semester)	Global-bewertung*	Lernzuwachs (%)	Modulrang nach...**	
			Global-bewertung	Lern-zuwachs
1 (1. Semester)	2,09	67,7%	7	7
2 (1. Semester)	2,70	63,0%	14	12
3 (1. Semester)	2,82	56,7%	17	14
4 (2. Semester)	1,84	72,3%	3	1
5 (2. Semester)	2,35	55,5%	**12**	**18**
6 (2. Semester)	2,06	56,6%	**6**	**15**
7 (2. Semester)	3,67	64,5%	**21**	**9**
8 (3. Semester)	1,59	71,2%	2	2
9 (3. Semester)	3,34	56,5%	20	16
10 (3. Semester)	2,13	69,5%	8	6
11 (4. Semester)	2,73	64,1%	15	10
12 (4. Semester)	2,70	49,2%	**13**	**19**
13 (4. Semester)	2,02	70,6%	5	5
14 (4. Semester)	2,23	64,9%	9	8
15 (5. Semester)	1,99	70,6%	4	4
16 (5. Semester)	2,24	63,2%	10	11
17 (5. Semester)	2,32	70,9%	**11**	**3**
18 (6. Semester)	2,73	56,8%	16	13
19 (6. Semester)	1,56	55,6%	**1**	**17**
20 (6. Semester)	2,88	11,5%	18	21
21 (6. Semester)	3,19	44,5%	19	20

* Schulnotenskala; ** Differenzen von mehr als 5 Rangplätzen sind fettgedruckt

von 64,3% (Platz 9) erzielt wurde. In diesem Zusammenhang fiel auf, dass die Studierenden das Modul bereits in der Motivationsabfrage negativ beurteilt hatten (Wichtigkeit: 3,64; Vorfreude: 4,20; Ruf des Moduls: 4,59).

Korrelationen zwischen den verschiedenen Evaluationsparametern

Die Ergebnisse der Korrelationsanalysen sind in Tabelle 2 dargestellt. Studentische Erwartungen *vor* Modulbeginn und die allgemeine Bewertung organisatorischer und struktureller Aspekte der Lehre *nach* Teilnahme am Modul korrelierten stark positiv. Auch zwischen der globalen Modulbewertung und der Zustimmung zur Aussage „Bezogen auf meine berufliche Zukunft schätze ich meinen Lernzuwachs in diesem Modul als sehr hoch ein“ fand sich eine starke positive Korrelation (r=0,94; p<0,001). Demgegenüber wiesen die Daten zum Lernerfolg nach VSE keine signifikante Korrelation mit den Erwartungen vor Modulbeginn auf. Zwischen dem Lernerfolg und den Ergebnissen der traditionellen Evaluation fand sich nur für zwei Aspekte eine schwache Korrelation (Varianzaufklärung maximal 22%). Der Lernerfolg korrelierte schwach mit der subjektiven Einschätzung, viel gelernt zu haben (r=0,44; p=0,044; Varianzaufklärung 19%) und nicht-signifikant mit der allgemeinen Modulbewertung (r=-0,42; p=0,061; Varianzaufklärung 18%; siehe Abbildung 2).

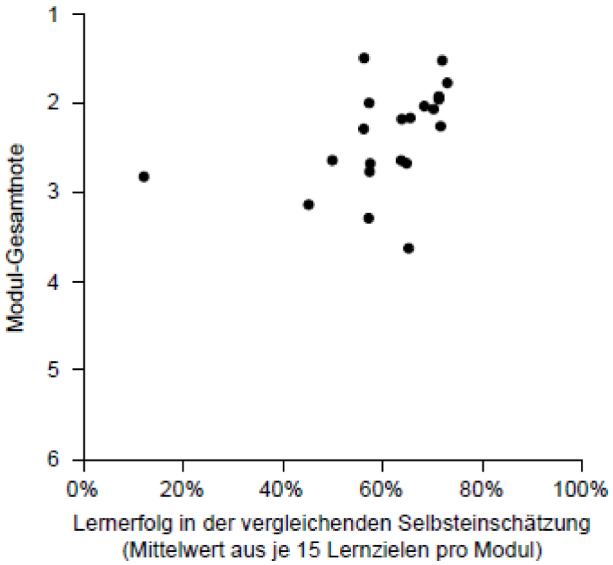

Abbildung 2: Korrelation zwischen Modul-Gesamtnote und berechnetem Lernerfolg für alle 21 Module des klinischen Studienabschnitts (r = -0,42; r^2 = 18%; p = 0,061).

Diskussion

Wichtigste Ergebnisse

In der vorliegenden Studie korrelierten globale Bewertungen von Lehrveranstaltungen signifikant mit der studen-

Tabelle 2: Korrelationen zwischen studentischen Bewertungen der 21 Module und dem mittels vergleichender Selbsteinschätzung berechneten Lernerfolg (pro Modul über 15 Lernziele gemittelt).

	Zeitpunkt der Datenerhebung										
	vor dem Modul				nach dem Modul						
	Wichtigkeit	Vorfreude	Ruf des Moduls		Inter-disziplinarität	Selbstlern-kompetenz	Subjektiver Lernzuwachs	Modulstruktur	Beibehaltung	Global-bewertung	Lernzuwachs (VSE)
Wichtigkeit	—	0,93**	0,83**		0,70**	0,67**	0,83**	0,75**	0,69**	0,71**	-0,23
Vorfreude	0,93**	—	0,89**		0,70**	0,59*	0,81**	0,70**	0,69**	0,76**	-0,3
Ruf des Moduls	0,83**	0,89**	—		0,85**	0,72**	0,85**	0,73**	0,80**	0,87**	-0,22
Interdisziplinarität	0,70**	0,70**	0,85**		—	0,88**	0,90**	0,85**	0,96**	0,95**	-0,34
Selbstlernkompetenz	0,67**	0,59*	0,72**		0,88**	—	0,88**	0,74**	0,86**	0,87**	-0,47*
Subjektiver Lernzuwachs	0,83**	0,81**	0,85**		0,90**	0,88**	—	0,85**	0,90**	0,94**	-0,44*
Modulstruktur	0,75**	0,70**	0,73**		0,85**	0,74**	0,85**	—	0,93**	0,84**	-0,29
Beibehaltung	0,69**	0,69**	0,80**		0,96**	0,86**	0,90**	0,93**	—	0,94**	-0,34
Globalbewertung	0,71**	0,76**	0,87**		0,95**	0,87**	0,94**	0,84**	0,94**	—	-0,42
Lernzuwachs (VSE)	-0,23	-0,3	-0,22		-0,34	-0,47*	-0,44*	-0,29	-0,34	-0,42	—

Legende: Angegeben sind Korrelationskoeffizienten nach Pearson; *p < 0,05: **p < 0,001; VSE = vergleichende Selbsteinschätzung.

tischen Erwartung vor Teilnahme an einen Modul sowie den retrospektiven Bewertungen von curricularer Struktur und subjektivem Lernzuwachs. Dagegen bestand keine Korrelation zwischen dem nach VSE berechneten Lernzuwachs und den studentischen Erwartungen vor Modulbeginn; zwischen dem berechneten Lernzuwachs und dem subjektiv wahrgenommenen Lernzuwachs fand sich eine schwache Korrelation. Entsprechend unterschied sich die Modul-Rangfolge – je nachdem, ob eine Globalbewertung oder der selbst eingeschätzte Lernzuwachs Maßstab war.

Stärken und Schwächen der Studie

Das hier vorgestellte, neue Evaluationsinstrument ermöglicht bei geringem Implementierungs-Aufwand erstmals auf der Ebene spezifischer Lernziele eine Beurteilung des Lernerfolgs durch vergleichende studentische Selbsteinschätzungen. Die hier vorgestellten Daten wurden im Rahmen des normalen Lehrbetriebs erhoben und stützen sich auf eine große Zahl von Einzelratings von über 500 Studierenden.

Die Validität *punktueller* Selbsteinschätzungen wurde in der Vergangenheit häufig kritisiert [5], [6], da verschiedene Konstrukt-irrelevante Faktoren die Genauigkeit von Selbsteinschätzungen beeinflussen können. Allerdings konnten Colthart et al. [7] zeigen, dass die Selbsteinschätzungsfähigkeit durch explizite Bewertungskriterien, Ankersetzung und Feedback verbessert werden kann. Somit stellt die Verwendung spezifischer Lernziele [8] eine entscheidende Voraussetzung für die Funktionalität des hier vorgestellten Instrumentes dar. Der Einfluss individueller Charakteristika auf das Ergebnis der Lernerfolgs-Evaluation nach VSE wird durch die wiederholte Erhebung von Selbsteinschätzungen in der gleichen Studierendengruppe reduziert.

Vergleich mit der Literatur und Bedeutung der Ergebnisse

Lehrqualität ist ein multidimensionales Konstrukt, in das unterschiedliche prozedurale, strukturelle, inhaltliche und ergebnisbezogene Parameter eingehen [9]. Studentische Evaluationen leisten einen Beitrag zur Bewertung der Lehrqualität an einer Fakultät. Allerdings unterliegen diese subjektiven Einschätzungen verschiedenen Einflüssen; daher ist nicht immer offensichtlich, welches Konstrukt im Detail in einer studentischen Evaluation abgebildet wird. Dies gilt insbesondere für globale Bewertungen einzelner Veranstaltungen, die zumeist mit Hilfe von Skalen nach dem Schulnotenprinzip erhoben werden. So zeigen die Daten der vorliegenden Studie, dass globale Bewertungen von Lehrveranstaltungen signifikant mit der studentischen Erwartung vor Teilnahme an einem Modul sowie den retrospektiven Bewertungen von curricularer Struktur und subjektivem Lernzuwachs (siehe Tabelle 2) korrelieren. Hier bieten sich zwei – sich nicht unbedingt widersprechende – Erklärungen an: Entweder den studentischen Bewertungen liegt tatsächlich ein Konstrukt von „guter Lehre“ zugrunde im Sinne eines gut strukturierten, interdisziplinären Unterrichts, in dem

gleichzeitig die Selbstlernkompetenz gefördert wird und der subjektive Lernzuwachs hoch ist.
Alternativ könnte vermutet werden, dass mit den unterschiedlichen Evaluationsparametern ein eigenes, homogenes Konstrukt abgebildet wird, innerhalb dessen eine Differenzierung der Einzelaspekte nicht mehr möglich ist. Im Extremfall wird lediglich die studentische Zufriedenheit mit der Lehre gemessen, die – wie an den hier beobachteten Korrelationen ersichtlich ist – stark mit den Erwartungen vor Modulbeginn korreliert. In der Tat bestätigen auch andere Untersuchungen, dass nicht nur strukturelle und prozedurale Aspekte [10], [11], [12], [13], [14], sondern auch das Auftreten der Dozenten [15], ihr Umgang mit den Studierenden [16] und der Ruf eines Dozenten bzw. einer Veranstaltung [17] sich auf studentische Globalbewertungen auswirken. Wenngleich ein professionelles Verhalten von Dozenten dem Konstrukt „gute Lehre“ hinzugerechnet werden kann, bleibt somit bei der Interpretation von studentischen Globalbewertungen meist unklar, welchen Beitrag dieser Aspekt zur Gesamtbewertung geleistet hat.
Neben strukturellen, prozeduralen und personenbezogenen Parametern sollte unbedingt auch der studentische Lernerfolg in die Bewertung von Lehrveranstaltungen einfließen [9]. Spontan wird man daran denken, hierfür die Ergebnisse fakultätsinterner Klausuren oder der Staatsexamina zur Beurteilung der Ergebnisqualität der Lehre heranzuziehen. Allerdings wurde erst kürzlich darauf hingewiesen, dass viele Prüfungen an deutschen medizinischen Fakultäten den internationalen Qualitätsstandards nicht genügen [18] und somit keinen Anspruch auf eine zufriedenstellende Validität erheben können. So repräsentieren Multiple Choice-Fragen der medizinischen Staatsexamina lediglich eine Dimension der ärztlichen Ausbildung (Wissen) und lassen keine Rückschlüsse auf die Qualität der praktischen Ausbildung zu. Zudem ist es insbesondere in reformierten Curricula kaum möglich, die aggregierten Examensergebnisse der Leistung einzelner Lehrveranstaltungen oder Fächer zuzuordnen, so dass eine intrafakultäre LOM-Verteilung auf dieser Grundlage verfehlt erscheint.

Praktische Implikationen

Bei der Interpretation von Evaluationsdaten ist generell darauf zu achten, dass das der Evaluation zugrunde liegende Konstrukt von „guter Lehre“ klar charakterisiert ist und dass die genutzten Erhebungsinstrumente dieses Konstrukt valide abbilden. Wünschenswert sind daher Evaluationsinstrumente, die valide spezifische Aspekte guter Lehre erfassen und eine möglichst geringe Kreuz-Korrelation zwischen unterschiedlichen Items aufweisen. Auf der Grundlage einer fehlenden bzw. geringen Korrelation zwischen Globalbewertungen und dem nach VSE bestimmten Lernzuwachs ist anzunehmen, dass das hier vorgestellte neue Evaluationsinstrument zusätzliche Informationen über den Lernerfolg liefert, die von traditionellen Evaluationsmethoden nicht erfasst werden. Ein Instrument mit diesen Eigenschaften empfiehlt sich für eine valide Unterrichts-Evaluation.

Zukünftiger Forschungsbedarf

Die studentische Definition des „Lernerfolgs“ wurde in der vorliegenden Arbeit nicht untersucht und sollte – insbesondere in Abgrenzung zur Definition durch die Lehrenden – Gegenstand künftiger Forschungsprojekte sein.
Für die zukünftige Verwendung unseres Verfahrens ist zu klären, ob eine Stichprobe von 15 Lernzielen pro Modul ausreichend ist, um das Spektrum einer Veranstaltung hinreichend abzubilden und somit den potentiellen Lernfortschritt verlässlich abzubilden. Zudem sollte ein eventueller Einfluss des Erhebungsinstruments [19] und der Rücklaufquote auf die Ergebnisse des neuen Evaluations-Instrumentes näher untersucht werden.
Aus der Notwendigkeit, an beiden Erhebungszeitpunkten Einschätzungen von den gleichen Individuen zu erhalten, ergeben sich mitunter datenschutzrechtliche Probleme, da die Evaluation medizinischer Lehrveranstaltungen anonym erfolgen sollte. Um dieses Problem zu umgehen, könnte anstatt einer Befragung zu zwei verschiedenen Zeitpunkten auch eine am Modulende liegende retrospektive Erhebung des subjektiven Leistungsstandes vor Modulbeginn erwogen werden [20]. Schließlich ist zu prüfen, inwieweit das neue Instrument auf andere Fakultäten und Curricula übertragbar ist.

Schlussfolgerung

Studentische Globalbewertungen von Veranstaltungen spiegeln nicht alle Aspekte der Lehrqualität wider und sind daher nur bedingt geeignet für die Zuteilung von LOM-Mitteln in der Lehre. Dagegen scheint das hier vorgestellte Instrument mit dem Lernzuwachs eine wichtige Dimension der Lehrqualität valide und unabhängig von gängigen Globalbewertungen abzubilden. Die Ergebnisse der Lernzuwachs-Evaluation nach VSE könnten daher zukünftig ein wesentlicher Bestandteil in Algorithmen zur Allokation von LOM-Mitteln sein.

Anmerkung

Die Autoren Raupach und Schiekirka teilen sich die Erstautorenschaft.

Interessenkonflikt

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

Literatur

1. Wissenschaftsrat. Empfehlungen zur Qualitätsverbesserung von Lehre und Studium. Berlin: Wissenschaftsrat; 2008.
2. 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. GMS Z Med Ausbild. 2007;24(2):Doc109. Zugänglich unter: http://www.egms.de/static/de/journals/zma/2007-24/zma000403.shtml
3. Müller-Hilke B. "Ruhm und Ehre" oder LOM für Lehre? - eine qualitative Analyse von Anreizverfahren für gute Lehre an Medizinischen Fakultäten in Deutschland. GMS Z Med Ausbild. 2010;27(3):Doc43. DOI: 10.3205/zma000680
4. Raupach T, Münscher C, Beißbarth T, Burckhardt G, Pukrop T. Towards outcome-based programme evaluation: Using student comparative self-assessments to determine teaching effectiveness. Med Teach. 2011;33(8):e446-453. DOI: 10.3109/0142159X.2011.586751
5. Falchikov N, Boud D. Student Self-Assessment in Higher Education: A Meta-Analysis. Rev Educ Res. 1989;59(4):395-430.
6. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. Jama. 2006;296(9):1094-1102. DOI: 10.1001/jama.296.9.1094
7. Colthart I, Bagnall G, Evans A, Allbutt H, Haig A, Illing J, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. Med Teach. 2008;30(2):124-145. DOI: 10.1080/01421590701881699
8. Harden RM. Learning outcomes as a tool to assess progression. Med Teach. 2007;29(7):678-682. DOI: 10.1080/01421590701729955
9. Rindermann H. Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. Z Evaluation. 2003(2):233-256.
10. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Intern Med. 2004;19(9):971-977. DOI: 10.1111/j.1525-1497.2004.40066.x
11. Kogan JR, Shea JA. Course evaluation in medical education. Teach Teach Educ. 2007;23(3):251-264. DOI: 10.1016/j.tate.2006.12.020
12. Marsh HW. The Influence of Student, Course, and Instructor Characteristics in Evaluations of University Teaching. Am Educ Res J. 1980;17(2):219-237.
13. Marsh HW. Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. J Educ Psychol. 1983;75:150-166. DOI: 10.1037/0022-0663.75.1.150
14. McKeachie W. Student ratings; the validity of use. Am Psychol. 1997;52(11):1218-1225. DOI: 10.1037/0003-066X.52.11.1218
15. 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):126-134. DOI: 10.1037/0022-0663.74.1.126
16. Jackson DL, Teal CR, Raines SJ, Nansel TR, Force RC, Burdsal CA. The dimensions of students' perceptions of teaching effectiveness. Educ Psychol Meas. 1999;59:580-596. DOI: 10.1177/00131649921970035
17. Griffin BW. Instructor Reputation and Student Ratings of Instruction. Contemp Educ Psychol. 2001;26(4). DOI: 10.1006/ceps.2000.1075
18. Möltner A, Duelli R, Resch F, Schultz JH, Jünger J. Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten. GMS Z Med Ausbild. 2010;27(3):Doc44. DOI: 10.3205/zma000681
19. Thompson BM, Rogers JC. Exploring the learning curve in medical education: using self-assessment as a measure of learning. Acad Med. 2008;83(10 Suppl):S86-S88. DOI: 10.1097/ACM.0b013e318183e5fd
20. Skeff KM, Stratos GA, Bergen MR. Evaluation of a Medical Faculty Development Program. Eval Health Prof. 1992;15(3):350-366. DOI: 10.1177/016327879201500307

Korrespondenzadresse:
PD Dr. med. Tobias Raupach, MME-D
Universitätsmedizin Göttingen, Abteilung Kardiologie & Pneumologie, 37099 Göttingen, Deutschland, Tel.: +49 (0)551/39-8922, Fax: +49 (0)551/39-6887
raupach@med.uni-goettingen.de

Bitte zitieren als
Raupach T, Schiekirka S, Münscher C, Beißbarth T, Himmel W, Burckhardt G, Pukrop T. Implementierung und Erprobung eines Lernziel-basierten Evaluationssystems im Studium der Humanmedizin. GMS Z Med Ausbild. 2012;29(3):Doc44. DOI: 10.3205/zma000814, URN: urn:nbn:de:0183-zma0008146

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2012-29/zma000814.shtml

Eingereicht: *08.07.2011*
Überarbeitet: *12.12.2011*
Angenommen: *12.01.2012*
Veröffentlicht: *15.05.2012*

Piloting an outcome-based programme evaluation tool in undergraduate medical education

Abstract

Aims: Different approaches to performance-oriented allocation of resources according to teaching quality are currently being discussed within German medical schools. The implementation of these programmes is impeded by a lack of valid criteria to measure teaching quality. An assessment of teaching quality should include structural and procedural aspects but focus on learning outcome itself. The aim of this study was to implement a novel, outcome-based evaluation tool within the clinical phase of a medical curriculum and address differences between the novel tool and traditional evaluation methods.

Methods: Student self-assessments before and after completion of a teaching module were used to compute performance gains for specific learning objectives. Mean performance gains in each module were compared to student expectations before the module and data derived from a traditional evaluation tool using overall course ratings at the end of the module.

Results: A ranking of the 21 modules according to computed performance gains yielded entirely different results than module rankings based on overall course ratings. There was no significant correlation between performance gain and overall ratings. However, the latter were significantly correlated to student expectations before entering the module as well as structural and procedural parameters (Pearson's r 0.7-0.9).

Conclusion: Performance gain computed from comparative self-assessments adds an important new dimension to course evaluation in medical education. In contrast to overall course ratings, the novel tool is less heavily confounded by construct-irrelevant factors. Thus, it appears to be more appropriate than overall course ratings in determining teaching quality and developing algorithms to guide performance-oriented resource allocation in medical education.

Keywords: evaluation, self-assessment, performance, learning objective, congruence, clinical studies

Tobias Raupach[1,2]
Sarah Schiekirka[3]
Christian Münscher[3]
Tim Beißbarth[4]
Wolfgang Himmel[5]
Gerhard Burckhardt[6]
Tobias Pukrop[7]

1 University Medical Centre Göttingen, Dept. of Cardiology & Pneumology, Göttingen, Germany

2 University College London, Department of Clinical, Educational and Health Psychology, London WC1E 7HB, United Kingdom

3 University Medical Centre Göttingen, Study Deanery, Göttingen, Germany

4 University Medical Centre Göttingen, Dept. of Medical Statistics, Göttingen, Germany

5 University Medical Centre Göttingen, Dept. of General Practice, Göttingen, Germany

6 University Medical Centre Göttingen, Dean for Study Affairs, Göttingen, Germany

7 University Medical Centre Göttingen, Dept. of Haematology and Oncology, Göttingen, Germany

Introduction

Following suggestions made by the German Council of Science and Humanities [1], half of all German medical schools internally allocated resources based on teaching performance in 2009. Such performance measures are also increasingly used to distribute funds to different medical schools within German federal states. In North Rhine-Westphalia, indicators used for this purpose can be attributed to three levels [2]: structural, procedural and outcome quality of teaching. Upon introduction of their algorithm, medical school Deans in North Rhine-Westphalia acknowledged the importance of structures (resources, time-tables) and processes (instructional format, examinations) but also suggested to include global performance indicators such as student results in high-stakes examinations, study time and drop-out rates

as measures of study outcome. Within the clinical phase of undergraduate medical education, these parameters cannot be broken down to clinical specialties and are thus of little use to guide performance-dependent resource allocation within a particular medical school. Consequently, there is a need to develop quality indicators referring to structures, processes and teaching outcome that can be used to rank courses or specialities.

A survey among German medical schools carried out in 2009 showed that many schools used evaluation data (i.e. student ratings) to guide internal fund distribution [3]. Traditional evaluation forms are predominantly made up of questions regarding structural and organisational aspects of teaching as well as global ratings on six-point scales resembling those used in German schools (with 1 being the best mark). While these instruments likely allow some appraisal of structures and processes, the extent to which they actually measure outcome quality is unknown.

The definition of desired outcomes in undergraduate medical education is a matter of debate; however, it is reasonable to suggest student performance gain as a possible primary outcome variable. Ideally, an evaluation tool would provide some measure of performance gain. Such a tool has been developed at Göttingen Medical School, and its reproducibility and *criterion* validity have recently been established [4]. By introducing 'learning success' as an outcome variable, it adds to traditional evaluation methods focusing on structural and procedural aspects of teaching quality.

In order to establish *discriminant* validity of the novel evaluation tool this study aimed at answering the following research questions:

1. Is there a difference between course rankings derived from performance gain data and traditional evaluation parameters? We hypothesized that the two approaches cover different aspects of teaching. As a consequence, rankings based on structural/procedural criteria should be considerably different from rankings based on performance gain.
2. Is there a correlation between student expectations towards specific courses, traditional evaluation parameters obtained at the end of courses and performance gain data? We hypothesized that traditional evaluation parameters would be highly correlated with each other. Since structural and procedural aspects considerably impact on teaching quality, we expected moderate correlations between the traditional and the novel evaluation tool.

Methods

Clinical curriculum and evaluation tools at Göttingen Medical School

The three-year clinical curriculum at our institution adopts a modular structure. There are 21 modules lasting two to seven weeks. The first clinical year covers basic practical skills, infectiology and pharmacotherapy, followed by 18 months of systematic teaching on the diagnosis and treatment of specific diseases. Aspects regarding differential diagnosis are focused on in the last 6 months.

1. Traditional evaluation tool: Students complete online evaluation forms (EvaSys®, Electric Paper, Lüneburg, Germany) at the end of each module. Ratings on organisational and structural aspects of teaching are obtained on six-point scales. The following statements are used for these ratings:
   1.1. "The implementation of interdisciplinary teaching was well done in this module." (Interdisciplinarity)
   1.2. "During this module, self-directed learning was promoted." (Self-directed learning)
   1.3. "Regarding my future professional life, I perceive my performance gain in this module as high." (Subjective performance gain)
   1.4. "I was very satisfied with the basic structure of this module (design, instructional formats, timetables)." (Structure)
   1.5. "This module should be continued unchanged." (Continuation)
2. The final item of this part of the questionnaire read "Please provide an overall school mark rating of the module."
3. In order to assess correlations between student expectations before a module and their ratings at the end of a module, an additional online survey at the beginning of each module was set up specifically for this study. In this survey, students were asked to rate the following statements on a six-point scale:
   3.1. "I believe that this module is important for my future professional life." (Importance)
   3.2. "I am looking forward to this module." (Anticipation)
   3.3. "The module has a good reputation with my more advanced fellow students." (Reputation)
4. Performance gain evaluation using comparative self-assessments (CSA): The novel tool used in this study calculates performance gain for specific learning objectives on the basis of repeated student self-assessments. At the beginning of each module, students are asked in an online survey to self-rate their performance levels regarding specific learning objectives (e.g. "I can interpret an electrocardiogram.") on a six-point scale anchored by "fully agree" (1) and "completely disagree" (6). Self-ratings were again obtained at the end of each module. The percent CSA gain for a specific learning objective was calculated by dividing the difference of mean values (pre-post) by the corrected mean of initial self-ratings across student cohorts according to the following formula (see figure 1):

$$\text{CSA gain [\%]} = \frac{\mu_{prä} - \mu_{post}}{\mu_{prä} - 1} \times 100$$

(μ_{pre} = mean initial self-assessment;
μ_{post} = mean self-assessment after the module)

Figure 1: CSA gain [%]

Aggregated CSA gain of a module was calculated as the mean of CSA gains obtained for 15 specific learning objectives which were derived from the Göttingen Catalogue of Specific Learning Objectives (http://www.med.uni-goettingen.de/de/media/G1-2_lehre/lernzielkatalog.pdf). The major difference between the aforementioned evaluation item "Regarding my future professional life, I perceive my performance gain in this module as high." and CSA gain was that subjective performance gain was obtained at one time-point only and was global in nature while CSA gain was calculated from ratings obtained at different time-points and related to specific learning objectives. A recently published study demonstrates a good correlation between CSA gain and an increase in objective performance measures [4].

Description of the study sample

In winter 2008/09, a total of 977 students were enrolled in clinical modules at our institution. All students were invited to participate in online evaluations via e-mail; each student automatically received three e-mails per survey containing a link to the online evaluation platform. Pre-surveys were opened three days before the beginning of a module and closed three days into the module. The same time-frame was used for post-surveys. Participation was voluntary, and all data were entered anonymously. Thus, further characterisation of the sample with regard to age and sex was not possible.

Data acquisition and analysis

Anonymous evaluation data collected during winter term 2008/09 were included in this study. During completion of the post-survey, students were asked to indicate whether they had also taken part in the pre-survey. Pre-post comparisons were only performed on data obtained from students who indicated to have completed both the pre- and the post-survey. Since data were aggregated across student cohorts, no individual labelling of students was necessary.

In order to address the first research question, the 21 modules were ranked according to mean global ratings or mean CSA gain values (aggregated from 15 specific learning objective gains). The second research question was addressed by calculating correlations between mean values of traditional evaluation items and aggregated CSA gain per module.

Analyses were run with SPSS® 14.0 (Illinois, USA). A Kolmogorov-Smirnov Test indicated that all data were normally distributed. Thus, correlations are reported as Pearson's r. Squared r values indicate the proportion of variance explained.

Results

Response rates

Of all 977 students enrolled in clinical modules, 573 provided a total of 51,915 ratings. Response rates in specific modules ranged from 36.7% to 75.4%.

Comparison of module rankings

Table 1 compares global student ratings and CSA gain values at the level of individual modules. Modules were ranked using both indicators (see the last two columns of the table). In some instances, ranks differed substantially depending on the method used. A difference of at least 6 ranks between the two methods was found for six out of 21 modules. For example, module no. 19 received the best overall rating (1.56) despite taking position 17 in the CSA gain ranking (55.6%). Conversely, module no. 7 ("Evidence-based medicine") received a mean overall rating of 3.67 and thus came last in this ranking while mean CSA gain was 64.3% (Rank 9). Notably, students had already provided negative ratings for this module in the pre-survey (Importance: 3.64; anticipation: 4.20; Reputation: 4.59).

Correlations between evaluation parameters

Results of correlation analyses are presented in Table 2. Student expectations *before* a module and ratings of organisational and structural aspects obtained *after* a module showed a strong positive correlation. There was also a strong correlation between overall ratings and approval of the statement "Regarding my future professional life, I perceive my performance gain in this module as high." ($r=0.94$; $p<0.001$). In contrast, there was no significant correlation between CSA gain and student expectations in the pre-survey. Weak correlations were observed between CSA gain and two variables included in the traditional evaluation form (maximum proportion of variance explained: 22%). CSA gain was weakly correlated with the subjective perception of having learned a lot ($r=-0.44$; $p=0.044$; proportion of variance explained 19%) and non-significantly with overall module ratings ($r=-0.42$; $p=0.061$; proportion of variance explained 18%; see Figure 2).

Table 1: Overall ratings, mean CSA gain and rankings of all 21 clinical modules, listed in their order of appearance in the three-year clinical curriculum. CSA, comparative self-assessment. Each term lasts 6 months.

Module number (term)	Overall rating*	CSA gain (%)	Module ranking by ** Overall rating	CSA gain
1 (Term 1)	2.09	67.7%	7	7
2 (Term 1)	2.70	63.0%	14	12
3 (Term 1)	2.82	56.7%	17	14
4 (Term 2)	1.84	72.3%	3	1
5 (Term 2)	2.35	55.5%	**12**	**18**
6 (Term 2)	2.06	56.6%	**6**	**15**
7 (Term 2)	3.67	64.5%	**21**	**9**
8 (Term 3)	1.59	71.2%	2	2
9 (Term 3)	3.34	56.5%	20	16
10 (Term 3)	2.13	69.5%	8	6
11 (Term 4)	2.73	64.1%	15	10
12 (Term 4)	2.70	49.2%	**13**	**19**
13 (Term 4)	2.02	70.6%	5	5
14 (Term 4)	2.23	64.9%	9	8
15 (Term 5)	1.99	70.6%	4	4
16 (Term 5)	2.24	63.2%	10	11
17 (Term 5)	2.32	70.9%	**11**	**3**
18 (Term 6)	2.73	56.8%	16	13
19 (Term 6)	1.56	55.6%	**1**	**17**
20 (Term 6)	2.88	11.5%	18	21
21 (Term 6)	3.19	44.5%	19	20

* six-point scale with 1 being the best option; ** Differences of >5 ranks are printed in bold letters

Table 2: Correlations between student ratings of the 21 modules and CSA gain (aggregated across 15 learning objectives per module).

	Time-point of data collection										
	before the module					**after the module**					
	Importance	Anticipation	Reputation		Inter-disciplinarity	Self-directed learning	Subjective performance gain	Structure	Continuation	Overall rating	CSA gain
Importance	—	0.93**	0.83**		0.70**	0.67**	0.83**	0.75**	0.69**	0.71**	-0.23
Anticipation	0.93**	—	0.89**		0.70**	0.59*	0.81**	0.70**	0.69**	0.76**	-0.3
Reputation	0.83**	0.89**	—		0.85**	0.72**	0.85**	0.73**	0.80**	0.87**	-0.22
Interdisciplinarity	0.70**	0.70**	0.85**		—	0.88**	0.90**	0.85**	0.96**	0.95**	-0.34
Self-directed learning	0.67**	0.59*	0.72**		0.88**	—	0.88**	0.74**	0.86**	0.87**	-0.47*
Subjective performance gain	0.83**	0.81**	0.85**		0.90**	0.88**	—	0.85**	0.90**	0.94**	-0.44*
Structure	0.75**	0.70**	0.73**		0.85**	0.74**	0.85**	—	0.93**	0.84**	-0.29
Continuation	0.69**	0.69**	0.80**		0.96**	0.86**	0.90**	0.93**	—	0.94**	-0.34
Overall rating	0.71**	0.76**	0.87**		0.95**	0.87**	0.94**	0.84**	0.94**	—	-0.42
CSA gain	-0.23	-0.3	-0.22		-0.34	-0.47*	-0.44*	-0.29	-0.34	-0.42	—

Legend: Cells contain Pearson's correlation coefficients; *p < 0.05: **p < 0.001; CSA, comparative self-assessment

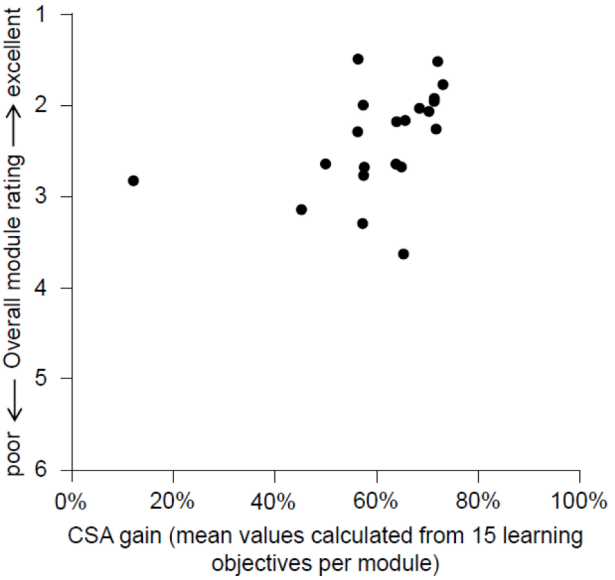

Figure 2: Correlation between overall module ratings and CSA gain for all 21 modules of the clinical phase of medical education (r = -0,42; r^2 = 18%; p = 0,061). CSA, comparative self-assessment

Discussion

Principal findings

In this study, we observed significant correlations between overall module ratings and student expectations before a module as well as with retrospective ratings of curricular structure and subjective performance gain. Conversely, there was no correlation between CSA gain and student expectations and only a weak correlation between CSA gain and perceived subjective performance gain. Accordingly, module rankings were different depending on them being based on overall ratings or CSA gain values.

Strengths and limitations

The new evaluation tool presented here facilitates a critical appraisal of performance gain by using comparative student self-assessments. It is easy to implement and addresses specific learning objectives. The data used for this study were obtained as part of our evaluation routine and comprised a large number of single ratings provided by over 500 students.

The validity of singular self-assessments has repeatedly been criticised in the past [5], [6] as their accuracy can be influenced by factors which are irrelevant to the construct examined. However, Colthart et al. [7] demonstrated that the ability to self-assess can be improved by using well-defined criteria, anchoring of scales and feedback. Accordingly, using specific learning objectives [8] for self-assessments is a prerequisite for the new tool to be functional. The potential impact of individual characteristics on CSA gain values is reduced by repeatedly collecting self-assessments from the same student group.

Relation to published research and significance of findings

Teaching quality is a multi-dimensional construct including several parameters related to process, structure, content and outcome of teaching [9]. Evaluation data provided by students contribute to the critical appraisal of teaching quality within medical schools. However, subjective ratings are influenced by various variables. As a consequence, details of the construct underlying student evaluation data may not be readily discernible. This particularly pertains to overall course ratings which are usually captured using a simple marking system. This study demonstrates that overall module ratings are strongly correlated with student expectations before taking a module as well as retrospective ratings of curricular structure and subjective performance gain (see Table 2). There are two possible explanations for this finding which may not be mutually exclusive: Student ratings may either feed into a construct of 'good teaching' which includes well-structured, interdisciplinary teaching that fosters self-directed learning and produces a considerable performance gain. Alternatively, it may be suggested that all seemingly different parameters used in traditional evaluation forms feed into the same, homogeneous construct within which specific aspects cannot be differentiated. In an extreme case, this would mean measuring mere student satisfaction which – as indicated by the correlations seen in this study – is strongly associated with student expectations before a module.

In fact, previous research indicates that overall ratings are not only influenced by structural and procedural aspects [10], [11], [12], [13], [14] but also by the behaviour of faculty [15], their rapport with students [16] and a lecturer's or a course's reputation [17]. While professional bearing of faculty might be attributed to a construct of 'good teaching', interpretation of student overall ratings is complicated by the extent of such contributions to the composite mark being largely unknown.

In addition to addressing structural and procedural parameters, a critical appraisal of teaching should take student learning outcome into consideration [9]. At first glance, student achievements in end-of-course assessments or high-stakes examinations might appear helpful in this regard. However, it has recently been reported that the assessments performed at many German medical schools do not live up to international quality standards [18] and thus cannot be regarded as being sufficiently valid. In addition, multiple choice questions address but one dimension of physician training (knowledge) and do not allow any conclusions to be drawn on the quality of practical training. Attributing exam results to teaching quality in specific courses or even specialties can be particularly difficult in reformed curricula. Thus, internal fund redistribution based on examination results does not seem advisable.

Practical implications

A clear definition of the construct of 'good teaching' and proof of validity for the tools used to assess are prerequisites for the interpretation of evaluation data. Accordingly, evaluation tools capturing specific aspects of good teaching and avoiding cross-correlations between different items are desirable. As correlations between overall ratings and CSA gain were either absent or weak in this study, the novel evaluation tool described here is likely to produce additional information that would have been missed by traditional evaluation tools. It might thus be used for a valid appraisal of teaching quality.

Research agenda

We did not assess students' definitions of performance gain in this study; this should be done in future studies, including research into teachers' definitions of learning outcome which may be different from students' views.
Future use of the novel evaluation tool needs to be informed by research regarding the number of learning objectives that need to be included per module in order to cover the whole range of aspects addressed in a module and, thus, produce reliable results. In addition, potential biases arising from the online data acquisition tool [19] and differing response rates across modules and their influence on CSA gain data needs to be studied in more detail.
The novel tool necessitates data collection from identical student groups at different time-points, thus potentially raising concerns regarding data protection issues as students should be able to complete evaluation forms anonymously. In order to solve this problem, the pre-/post-design could be changed to a singular data collection period at the end of each module during which students would provide retrospective ratings of their performance levels at the beginning of the module [20]. Finally, future research needs to show to what extent the novel tool can be transferred to other medical schools and curricula.

Conclusions

Overall course ratings provided by students do not address all aspects of teaching quality and are of limited use to guide fund redistribution within medical schools. In contrast, by providing an estimate of actual performance gain, the tool described here appears to cover an important dimension of teaching quality. It was found to be valid and independent of traditional global ratings. Therefore, CSA gain might become an integral part of algorithms to inform performance-guided resource allocation within medical schools.

Note

The authors Raupach and Schiekirka have equally contributed to this manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Wissenschaftsrat. Empfehlungen zur Qualitätsverbesserung von Lehre und Studium. Berlin: Wissenschaftsrat; 2008.
2. 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. GMS Z Med Ausbild. 2007;24(2):Doc109. Zugänglich unter: http://www.egms.de/static/de/journals/zma/2007-24/zma000403.shtml
3. Müller-Hilke B. "Ruhm und Ehre" oder LOM für Lehre? - eine qualitative Analyse von Anreizverfahren für gute Lehre an Medizinischen Fakultäten in Deutschland. GMS Z Med Ausbild. 2010;27(3):Doc43. DOI: 10.3205/zma000680
4. Raupach T, Münscher C, Beißbarth T, Burckhardt G, Pukrop T. Towards outcome-based programme evaluation: Using student comparative self-assessments to determine teaching effectiveness. Med Teach. 2011;33(8):e446-453. DOI: 10.3109/0142159X.2011.586751
5. Falchikov N, Boud D. Student Self-Assessment in Higher Education: A Meta-Analysis. Rev Educ Res. 1989;59(4):395-430.
6. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. Jama. 2006;296(9):1094-1102. DOI: 10.1001/jama.296.9.1094
7. Colthart I, Bagnall G, Evans A, Allbutt H, Haig A, Illing J, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. Med Teach. 2008;30(2):124-145. DOI: 10.1080/01421590701881699
8. Harden RM. Learning outcomes as a tool to assess progression. Med Teach. 2007;29(7):678-682. DOI: 10.1080/01421590701729955
9. Rindermann H. Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. Z Evaluation. 2003(2):233-256.
10. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Intern Med. 2004;19(9):971-977. DOI: 10.1111/j.1525-1497.2004.40066.x
11. Kogan JR, Shea JA. Course evaluation in medical education. Teach Teach Educ. 2007;23(3):251-264. DOI: 10.1016/j.tate.2006.12.020
12. Marsh HW. The Influence of Student, Course, and Instructor Characteristics in Evaluations of University Teaching. Am Educ Res J. 1980;17(2):219-237.
13. Marsh HW. Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. J Educ Psychol. 1983;75:150-166. DOI: 10.1037/0022-0663.75.1.150

14. McKeachie W. Student ratings; the validity of use. Am Psychol. 1997;52(11):1218-1225. DOI: 10.1037/0003-066X.52.11.1218

15. 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):126-134. DOI: 10.1037/0022-0663.74.1.126

16. Jackson DL, Teal CR, Raines SJ, Nansel TR, Force RC, Burdsal CA. The dimensions of students' perceptions of teaching effectiveness. Educ Psychol Meas. 1999;59:580-596. DOI: 10.1177/00131649921970035

17. Griffin BW. Instructor Reputation and Student Ratings of Instruction. Contemp Educ Psychol. 2001;26(4). DOI: 10.1006/ceps.2000.1075

18. Möltner A, Duelli R, Resch F, Schultz JH, Jünger J. Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten. GMS Z Med Ausbild. 2010;27(3):Doc44. DOI: 10.3205/zma000681

19. Thompson BM, Rogers JC. Exploring the learning curve in medical education: using self-assessment as a measure of learning. Acad Med. 2008;83(10 Suppl):S86-S88. DOI: 10.1097/ACM.0b013e318183e5fd

20. Skeff KM, Stratos GA, Bergen MR. Evaluation of a Medical Faculty Development Program. Eval Health Prof. 1992;15(3):350-366. DOI: 10.1177/016327879201500307

Corresponding author:
PD Dr. med. Tobias Raupach, MME-D
University Medical Centre Göttingen, Dept. of Cardiology & Pneumology, 37099 Göttingen, Germany, Phone +49 (0)551/39-8922, Fax: +49 (0)551/39-6887
raupach@med.uni-goettingen.de

Please cite as
Raupach T, Schiekirka S, Münscher C, Beißbarth T, Himmel W, Burckhardt G, Pukrop T. Implementierung und Erprobung eines Lernziel-basierten Evaluationssystems im Studium der Humanmedizin. GMS Z Med Ausbild. 2012;29(3):Doc44. DOI: 10.3205/zma000814, URN: urn:nbn:de:0183-zma0008146

This article is freely available from
http://www.egms.de/en/journals/zma/2012-29/zma000814.shtml

Received: *2011-07-08*
Revised: *2011-12-12*
Accepted: *2012-01-12*
Published: *2012-05-15*