Leitsymptomvorlesungen im klinischen Studienabschnitt - Effekte evaluationsbasierter Interventionen auf eine Großgruppen-Lehrveranstaltung

Olaf Kuhnigk[1,2]
Katja Weidtmann[2]
Sven Anders[3]
Bernd Hüneke[4]
René Santer[5]
Sigrid Harendza[6]

1 Universitätsklinikum Hamburg-Eppendorf, Klinik für Psychiatrie und Psychotherapie, Hamburg, Deutschland
2 Universitätsklinikum Hamburg-Eppendorf, Prodekanat für Lehre, Hamburg, Deutschland
3 Universitätsklinikum Hamburg-Eppendorf, Institut für Rechtsmedizin, Hamburg, Deutschland
4 Universitätsklinikum Hamburg-Eppendorf, Klinik und Poliklinik für Geburtshilfe und Pränatalmedizin, Hamburg, Deutschland
5 Universitätsklinikum Hamburg-Eppendorf, Klinik und Poliklinik für Kinder- und Jugendmedizin, Hamburg, Deutschland
6 Universitätsklinikum Hamburg-Eppendorf, III. Medizinische Klinik, Hamburg, Deutschland

Zusammenfassung

In der medizinischen Hochschulausbildung ist das Veranstaltungsformat „Vorlesung“ trotz vielfacher Kritik weiterhin ein wesentliches Element bestehender Curricula. Im Rahmen der Studienreform an der Medizinischen Fakultät der Universität Hamburg im Jahr 2004 wurden im reformierten klinischen Curriculum die fachspezifischen Vorlesungen durch Leitsymptomvorlesungen (LSV) ersetzt, die sich durch alle sechs Themenblöcke der Studienjahre drei bis fünf ziehen. Da die regelmäßigen Trimesterabschlussevaluationen der Studierenden auf einen Verbesserungsbedarf der LSV hindeuteten, wurde in dieser Studie die LSV mit Terminevaluationen durch anwesende Studierende und durch geschulte Auditoren (PJ-Studierende und wissenschaftliche Mitarbeiter) untersucht. Auf der Basis dieser Ergebnisse erfolgte ein schriftliches Feedback der Evaluationsdaten an die Lehrenden in Kombination mit Informationsmaterial über eine optimale Gestaltung der LSV nach modernen didaktischen Gesichtspunkten. In einer zweiten Erhebung wurden die Effekte der Intervention untersucht. Es zeigte sich, dass auf der Ebene von Schulnoten nach der Intervention nur geringe Verbesserungen der Qualität der LSV bemerkbar waren. In der Bewertung der Einzelmerkmale, insbesondere zur didaktischen Qualität, ließen sich hingegen signifikante Verbesserungen aufzeigen. Insgesamt bewerteten Studierende die LSV in der ersten Erhebungsphase bezogen auf die Einzelmerkmale signifikant positiver als die geschulten Auditoren. Dieser Effekt war in der zweiten Erhebungsphase nicht mehr nachweisbar. Unter den Auditoren bestand eine gute Inter-Rater-Reliabilität. Durch diese Untersuchung wurde insbesondere deutlich, dass die Einbettung von Lehrveranstaltungen auf struktureller und personeller Ebene in curriculare Gesamtkonzepte regelmäßig durch Qualitätssicherungsmaßnahmen zu begleiten ist. Wie häufig und in welcher Form strukturierte Rückmeldungen nach Evaluationen erfolgen müssen, um nachhaltige Effekte auf die didaktische Qualität der Lehrveranstaltungen zu haben, müssen zukünftige Studien zeigen.

Schlüsselwörter: Vorlesung, Evaluation, Audit, Intervention, Leitfaden, didaktische Fertigkeiten, Dozententraining, Qualitätssicherung

Einleitung

Vorlesungen als Lernformat

Das Veranstaltungsformat „Vorlesung“ ist trotz Kritik und sich daraus ergebender Studienreformen weiterhin ein für die medizinische Ausbildung wichtiges didaktisches Element [6]. Einerseits wird am traditionell-systematischen Frontalvortrag kritisiert, dass er für die Entwicklung von eigenständigem Denken nicht förderlich sei. Andererseits bieten Vorlesungen die Möglichkeit, Gruppen von Lernenden Informationen ökonomisch und ressourceneffizient zu vermitteln, einen Einstieg in komplexere Themen zu liefern, sowie aktuelle Forschungsergebnisse und persönliche, klinische oder wissenschaftliche Erfahrungen darzustellen [5]. Damit diese potentiellen Vorteile der

Vorlesung genutzt werden können, sollte das Veranstaltungsformat „Vorlesung“ in die an Lernzielen orientierten curricularen Rahmenbedingungen eingefügt [15] und zur Stimulation des eigenverantwortlichen Lernens der Studierenden mit anderen Lernformaten verknüpft werden [13]. Inhaltlich hat sich hierbei vor allem das fallbezogene Unterrichtsformat bewährt [10].

Leitsymptomvorlesungen im Hamburger Curriculum

An der Medizinischen Fakultät der Universität Hamburg erfolgte im Jahr 2004 eine umfassende Reform des klinischen Studienabschnitts, wobei das reformierte Klinische Curriculum Medizin (KliniCuM) auf eine fächerübergreifende und praxisbezogene Ausbildung abzielt [30]. Der Unterricht der Studienjahre drei bis fünf ist in sechs Themenblöcke und ein Wahlfach aufgeteilt und am Hamburger Lernzielkatalog [29] orientiert, der die verschiedenen Lerndimensionen und Kompetenzebenen abbildet. Systematische, fachspezifische Vorlesungen wurden im Zuge dieses Reformprozesses abgeschafft und durch Vorlesungen ersetzt, die sich fallbezogen an führenden Symptomen verschiedener Krankheiten orientieren. Das Konzept der Leitsymptomvorlesung (LSV) ist ein zentraler Bestandteil des KliniCuM, der sich als roter Faden durch alle Themenblöcke zieht und Zusammenhänge zwischen Inhalten anderer Veranstaltungen (z.B. problemorientierte Tutorien, Unterricht am Krankenbett) erkennbar macht. Erste Ergebnisse der Besuchsquoten und der Evaluationen deuteten auf eine größere Studierendenzufriedenheit mit dem neuen Format im Vergleich zu Vorlesungen vor der Studienreform hin. Allerdings zeigte sich auch früh Verbesserungsbedarf, wobei jedoch die konkreten Kritikpunkte der Studierenden an der LSV in eher allgemein gehaltenen Kommentaren der nach dem jeweiligen Trimester durchgeführten Abschlussevaluation weitgehend unklar blieben [32].

Evaluation als Interventionsbasis

Studierende sind in der Lage, die didaktische Qualität von Lehrveranstaltungen reliabel und valide zu bewerten [12], [22]. Gleichzeitig wird jedoch gefordert, Lehrevaluationen nicht allein auf studentische Beurteilungen zu stützen [11], [21]. Da die LSV nach den oben genannten Kriterien [10], [13] für eine inhaltlich sinnvolle Nutzung dieses Lernformats konzipiert worden war, führten wir zur Qualitätskontrolle die hier vorliegende Studie durch. Diese beinhaltet eine detaillierte Untersuchung der an der LSV empfundenen Mängel und die Beobachtung der Auswirkungen einer auf dieser Mängelanalyse basierenden Intervention zur Optimierung dieser Veranstaltungsform.

Methoden

Fragestellung und Hypothesen

In der vorliegenden Studie wurden zwei Fragen untersucht. Erstens: Lassen sich bei Terminevaluationen durch Vorlesungsteilnehmende und Audits durch geschulte Auditoren nach einer Intervention, die auf dem Boden der erhobenen Ergebnisse durchgeführt wird, Veränderungen in der Bewertung der LSV feststellen? Zweitens: Unterscheiden sich die Bewertungen der anwesenden Studierenden von den Urteilen geschulter Auditoren?
Die zentralen Hypothesen der Studie lauteten:

1. Die LSV wird nach einer Intervention, insbesondere in den didaktischen Bewertungen, positiver beurteilt als vor der Intervention.
2. Die Bewertungen geschulter Auditoren sind einheitlicher und insgesamt kritischer als die der Studierenden.

Erhebungsinstrumente

Eine Übersicht über den Ablauf der Studie, die sich in zwei Erhebungsphasen und eine dazwischen liegende Interventionsphase gliedert, ist in Abbildung 1 dargestellt.
In einer Pilotphase war eine Checkliste für Audits der Leitsymptomvorlesung konzipiert und validiert worden [32]. Sie beinhaltete sieben Merkmale zu Struktur und Inhalt der Vorlesung sowie neun Merkmale zu didaktischen Fertigkeiten der Dozierenden. Die Gruppe der Auditoren setzte sich aus acht wissenschaftlichen Mitarbeitern und 14 Studierenden im Praktischen Jahr zusammen. Es wurden jeweils Auditorenpaare aus einem wissenschaftlichen Mitarbeiter und einem Studierenden ausgelost, um eventuelle systematische Unterschiede, z.B. durch den Status (Nicht-Studierender/Studierender) begründete divergierende Perspektiven, zu kontrollieren. Vor der Pilotphase erfolgte eine dreistündige Schulung aller Auditoren, in der das Testinstrument erklärt und ein für die Evaluation standardisiertes methodisches Vorgehen eingeübt wurden.
Außerdem wurde ein Fragebogen zur Terminevaluation für die in der LSV anwesenden Studierenden entworfen. Dieser enthielt sowohl zentrale Aspekte der Vorlesung wie Orientierung an Leitsymptomen, Praxisbezug oder strukturierter Aufbau, als auch Merkmale der Lehrperson, z.B. Art des Umgangs mit den Studierenden, Verständlichkeit und Anschaulichkeit des Vortrags. Die Terminevaluation dokumentierte zudem Charakteristika der Studierenden, z.B. Geschlecht und Regelmäßigkeit des Vorlesungsbesuchs. Merkmale der Fragebögen und Checklisten waren auf einer 6-stufigen Likert-Skala zu bewerten (1: „trifft gar nicht zu“ bis 6: „trifft sehr zu“). Freitextkommentare waren darüber hinaus möglich und die Gesamtbewertung der Veranstaltung erfolgte in Form einer Schulnote (1=sehr gut, 2=gut, 3=befriedigend, 4=ausreichend, 5=mangelhaft, 6=ungenügend). Die Datenerhebungsinstrumente erwiesen sich in den Pilottestungen als prakti-

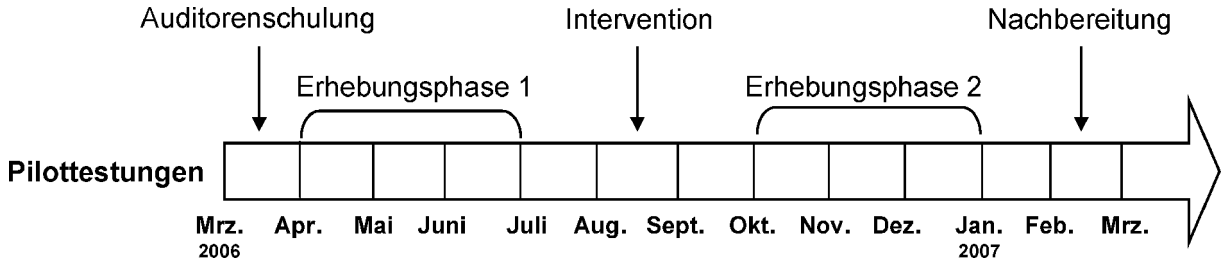

Abbildung 1: Zeitlicher Ablauf der Studie mit Übersicht über die einzelnen Arbeitsschritte

kabel und verständlich. Sie wurden für die Erhebungen nur noch geringfügig modifiziert, zum Beispiel durch Umstellung der Fragenreihenfolge oder durch Verdeutlichung von Merkmalen durch Nennung von Beispielen in Klammern.

Da sich in den Pilottestungen in einigen Bereichen deutliche Unterschiede zwischen den beiden Auditorengruppen (wissenschaftliche Mitarbeiter / PJ-Studierende) ergaben, wurde vermutet, dass die initiale Schulung der Auditoren nicht alle wesentlichen Aspekte ausreichend hatte vermitteln können. Es erfolgte daher eine zweite Schulung vor Beginn der ersten Erhebungsphase mit einer erläuternden Zusammenfassung des Konzepts der LSV im KliniCuM, einer Illustration der Orientierung an Leitsymptomen und von Bezügen zwischen den Fächern des Themenblocks sowie der Vermittlung von leitsymptomorientiertem Fachwissen anhand von konkreten Beispielen. Die Übereinstimmung der Bewertungen zwischen den beiden Auditorengruppen wurde mittels Berechnung des Intraklassenkoeffizienten ermittelt [34].

Design und Stichprobe

Alle an der LSV teilnehmenden Dozierenden wurden über die Durchführung dieser Evaluationsstudie informiert. Die Dozierenden wurden nicht darüber in Kenntnis gesetzt, ob ihre Vorlesung für eine Evaluation ausgelost wurde bzw. welcher der von ihnen unterrichteten Termine ausgelost wurde. Für die erste Erhebungsphase wurde aus den insgesamt 247 Einzelterminen der LSV aller sechs Themenblöcke im Trimester April bis Juli 2006 randomisiert eine repräsentative Stichprobe ausgewählt, die ca. ein Drittel der LSV pro Themenblock umfasste (insgesamt über alle Themenblöcke n=85). Dieses Vorgehen ist als Ziehung geschichteter Stichproben zu charakterisieren und wurde gewählt, da die Grundgesamtheit (sämtliche LSV aller Themenblöcke) als sehr heterogen einzuschätzen ist, also die Merkmalsausprägungen der Grundgesamtheit große Unterschiede aufweisen könnten. Um die bestehenden Schattierungen der Grundgesamtheit in einer Stichprobe ausreichend abzubilden, müsste nach dem Prinzip der reinen Zufallsauswahl die Stichprobe sehr groß sein, um die Repräsentativität zu gewährleisten. Zur Lösung dieser Problematik wurde die Grundgesamtheit in disjunkte Klassen (Schichten) aufgeteilt. Dabei wurde angenommen, dass sich die Elemente einer jeden Schicht bezüglich der untersuchten Frage ähnlich verhalten und dass sich dementsprechend Elemente aus unterschiedlichen Schichten durch die als relevant definierten verschiedenen Eigenschaften auszeichnen [8]. In dieser Untersuchung bildeten die sechs Themenblöcke die Schichten. Von proportional geschichteten Stichproben wurde aus jeder Klasse eine reine Zufallsstichprobe durch Auslosung der einzelnen Vorlesungstermine gezogen. Die Auswertung der Daten erfolgte mit dem statistischen Auswertungsprogramm SPSS 16.0. Mittelwertvergleiche wurden anhand des t-Tests für unabhängige Stichproben durchgeführt (Signifikanzniveau $p<0,05$). Die statistischen Tests wurden deskriptiv verwendet.

Basierend auf den Daten der ersten Erhebungsphase wurde folgende Intervention durchgeführt. Es wurden drei Gruppen ermittelt, die eine Intervention erhalten sollten: die in der Stichprobe evaluierten Lehrenden (Gruppe 1), alle an der LSV beteiligten Lehrenden und Abteilungsdirektoren, die bisher nicht evaluiert wurden (Gruppe 2), Studierende und die interessierte Öffentlichkeit (Gruppe 3). Komponenten der Intervention und ihre Zuordnung waren:

- Anschreiben: Alle Personen der Gruppen 1 und 2 erhielten ein persönliches Anschreiben, das über Hintergrund, Vorgehen und Ziele des Projekts sowie über die jeweiligen Feedback-Komponenten und die Ansprechpartner bei Rückfragen informierte.
- Allgemeines Feedback: Alle drei Gruppen erhielten eine Auswertung der Ergebnisse der ersten Erhebungsphase mit nicht personenbezogenen, allgemeinen Statistiken.
- Individuelles Feedback: Alle Personen der Gruppe 1 erhielten ihr personenbezogenes Feedback mit den Bewertungen und Freitextkommentaren der Auditoren und Studierenden.
- LSV-Leitfaden: Alle Gruppen erhielten einen auf den Daten der ersten Erhebung zusammengestellten Leitfaden als „Gold-Standard“ für die Ausgestaltung von LSV mit konkreten Tipps zu Inhalt und Form.
- Publikation: Gruppe 1 und 2 erhielten die Veröffentlichung „Teaching large groups“ [7], eine Publikation zur Ausgestaltung von Großgruppen-Lehrveranstaltungen in der medizinischen Ausbildung, die in sehr kompakter Darstellung einfach umzusetzende Anregungen für die Ausgestaltung von Vorlesungen nach modernen didaktischen Prinzipien enthält.
- Für alle Personen der Gruppe 3 wurden Informationen zur LSV und der LSV-Leitfaden auf die Homepage des Prodekanats für Lehre ins Internet gestellt.

Für die zweite Erhebungsphase wurden die in Erhebungsphase 1 evaluierten Vorlesungen (n=78, die geringere Zahl erklärt sich durch drei Ausfälle von Audits und vier

Ausfälle von Vorlesungen) auf der Basis ihrer Bewertung nach deutschen Schulnoten in drei Ligen eingeteilt (siehe Tabelle 1). Die Auswahl der zu evaluierenden LSV erfolgte als Ziehung von Quotenstichproben [4], also als bewusste Auswahl, die für die zu ziehenden Stichproben anstrebt, in der Grundgesamtheit vorliegende Strukturen nachzuahmen. Die Grundgesamtheit ist hier durch die sechs Themenblöcke und die drei Ligen definiert, wobei die Termine entsprechend repräsentativ ausgewählt wurden. Da in der Untersuchung insbesondere die Frage beantwortet werden sollte, ob sich die didaktische Qualität der LSV nach Durchführung der oben beschriebenen Intervention verbessert, wurde dies durch die studentische Evaluation der Termine per Schulnote kriterienbasiert operationalisiert. Unter Heranziehung von Überlegungen zum angestrebten minimalen Unterschied bei der studentischen Bewertung sowie zur Testpower wurde eine erforderliche Stichprobengröße von studentischen Bewertungen von n=633 ermittelt [20]. In diesem Fall waren folgende relevante Kriterien bestmöglich erfüllt: Effektstärke d=0,3 (d=0.1: kleiner Effekt, d=0,3: mittlerer Effekt, d=0,5: großer Effekt), minimaler Mittelwertsunterschied Δ=+0,255, Teststärke 1-ß=0,8 und α=0,05. Da in der ersten Erhebungsphase pro Vorlesungstermin durchschnittlich etwa n=35 studentische Bewertungen erhoben werden konnten, ergaben sich damit n=18 notwendige Terminevaluationen im Rahmen der zweiten Erhebungsphase. Hierbei handelte es sich in 14 Fällen um dieselben Lehrpersonen wie in Erhebungsphase 1. Die Themen der 18 Vorlesungen waren in allen Fällen dieselben wie in der ersten Erhebungsphase. Für eine gleichmäßige Verteilung dieser Messungen auf die Struktur der Grundgesamtheit ergab sich die Auswahl je eines Vorlesungstermins aus jedem der sechs Themenblöcke und jeder der drei Ligen.

Tabelle 1: Verteilung der Ergebnisse der LSV der Erhebungsphase 1 auf die drei Ligen für die Auswahl der Termine der Erhebungsphase 2. (Für die zweite Erhebungsphase wurden die in der ersten Erhebungsphase evaluierten Vorlesungen (n=78) auf der Basis ihrer Bewertung nach deutschen Schulnoten in drei Ligen eingeteilt).

Liga	Bewertung	$M_{Schulnote}$	LSV TB 1	LSV TB 2	LSV TB 3	LSV TB 4	LSV TB 5	LSV TB 6	$LSV_{ges.}$	$n_{Stud.}$
1	gut	M ≤ 2,0	3	6	5	5	2	4	25	1017
2	mittel	2,1 ≤ M ≤ 2,3	6	1	4	3	5	4	23	1161
3	schlecht	M ≥ 2,4	4	7	6	5	7	1	30	1176
LSV_{gesamt}			13	14	15	13	14	9	78	
$n_{Stud.}$			630	634	726	595	348	421		3354

M: Mittelwert, LSV: Leitsymptomvorlesung, TB: Themenblock, n: Anzahl

Ergebnisse

Veränderungen in der Bewertung der LSV nach Schulnoten

Auf Basis der Schulnoten zeigen bei der studentischen Terminevaluation von 18 erfassten LSV-Mittelwertvergleichen derselben Veranstaltungen in Erhebungsphase 1 und 2 fünf signifikante Verbesserungen (28%), drei signifikante Verschlechterungen (17%) und zehn unveränderte Beurteilungen (55%) (siehe Tabelle 2). Damit zeigen die studentischen Evaluationen für die Mehrheit der in der zweiten Erhebung erfassten LSV Termine keine Veränderung. Das oben beschriebene Kriterium eines minimalen Mittelwertunterschieds bei der Schulnote von Δ=+0,255 wird bei acht Vorlesungen erfüllt (44%).

Die Bewertung der LSV anhand der Schulnotenskala durch die Auditoren ergibt ebenfalls kein einheitliches Bild bezüglich eines Effekts der Intervention (siehe Tabelle 3). Das prozentuale Verhältnis der Veränderungen entspricht dem oben dargestellten der studentischen Terminevaluation, wobei vier der fünf als verbessert bewerteten LSV übereinstimmen. In der ersten Erhebungsphase beurteilten in der Gruppe der Auditoren die wissenschaftlichen Mitarbeiter sechs, in der zweiten Erhebungsphase sieben von 18 Vorlesungen um eine Schulnote schlechter als die PJ-Studierenden.

Tabelle 3: Gegenüberstellung der von den Auditoren vergebenen Schulnoten aus den Erhebungsphasen 1 und 2 für die 18 erfassten Termine

Nr.	TB	$Auditor\text{-}Noten_{Erhebung1}$		$Auditor\text{-}Noten_{Erhebung2}$		Gegenüberstellung#
		PJler	Wiss.	PJler	Wiss.	
1	1	1	2	2	3	verschlechtert
2	5	3	4	3	4	unverändert
3	2	2	3	2	3	unverändert
4	3	2	2	4	4	verschlechtert
5	5	3	3	2	3	unverändert
6	6	4	5	2	1	**verbessert**
7	2	2	2	1	1	**verbessert**
8	6	2	2	2	2	unverändert
9	4	4	4	2	3	**verbessert**
10	2	2	2	2	1	unverändert
11	4	4	3	4	2	unverändert
12	1	3	3	2	3	unverändert
13	6	2	1	1	1	unverändert
14	1	3	4	2	2	**verbessert**
15	3	3	2	2	1	**verbessert**
16	4	1	2	1	2	unverändert
17	5	3	3	4	4	verschlechtert
18	3	2	2	2	2	unverändert

TB: Themenblock; Schulnoten: 1=sehr gut bis 6=ungenügend;
verbessert: Bei der zweiten Erhebung waren die Schulnoten von beiden Auditoren mindestens um eine Note verbessert oder die Note eines Auditors war mindestens zwei Noten besser als in der ersten Erhebung.
verschlechtert: Bei der zweiten Erhegung waren die Schulnoten von beiden Auditoren um mindestens eine Note verschlechtert oder die Note eines Auditors war um mindestens zwei Noten schlechter als in der ersten Erhebung.
In der Spalte „Gegenüberstellung" sind die in der zweiten Erhebungsphase besser bewerteten LSV fett markiert. Dies ist jedoch nur für fünf LSV der Fall.

Didaktische Bewertung nach Einzelmerkmalen

Ein differenzierteres Bild als die Schulnoten liefert die Gegenüberstellung der Bewertungen durch die Auditoren auf Ebene der Einzelmerkmale aus beiden Erhebungen (siehe Tabelle 4). Es ergeben sich sechs signifikant bessere Beurteilungen nach der Intervention in der zweiten Erhebungsphase und alle anderen zeigen bis auf drei einen positiven Trend. Insgesamt weisen die Verbesserungen bei den Merkmalen „Orientierung an Leitsymptomen", „Anregung zum Mitdenken", „Entsprechung des Konzepts LSV", „interaktive Gestaltung", „anschauliche Darstellung" und „Bemühung um Lernerfolg" hohe Effektstärken auf.

Tabelle 2: Vergleich der Schulnoten der Terminevaluationen aus den Erhebungsphasen 1 und 2

Nr.	TB	Erhebung 1			$Liga_{Erhebung1}$	Erhebung 2			$Liga_{Erhebung2}$	$\Delta_{Erhebung\ 1\text{-}2}$	Signifikanz*	Erfüllung Kriterium#
		$M_{Schulnote}$	SD	$n_{Stud.}$		$M_{Schulnote}$	SD	$n_{Stud.}$				
1	1	**2,2**	,75	67	mittel	**2,7**	,96	47	schlecht	**- ,5**	,001 (,6)	**nein**
2	5	**2,3**	,53	56	mittel	**2,4**	,87	56	schlecht	**- ,1**	,489 (,1)	**nein**
3	2	**2,4**	,86	70	schlecht	**2,0**	,76	50	gut	**+ ,4**	**,002 (,5)**	**ja**
4	3	**2,3**	,85	56	mittel	**2,2**	,79	49	mittel	**+ ,1**	,466 (,1)	**nein**
5	5	**2,0**	,59	22	gut	**2,1**	,72	54	mittel	**- ,1**	,442 (,1)	**nein**
6	6	**2,6**	,79	64	schlecht	**2,2**	,68	66	mittel	**+ ,4**	**,008 (,5)**	**ja**
7	2	**2,1**	,62	67	mittel	**1,9**	,61	78	gut	**+ ,2**	**,004 (,2)**	**ja**
8	6	**2,2**	1,06	56	mittel	**1,9**	,80	60	gut	**+ ,3**	,076 (,4)	**ja**
9	4	**2,8**	,84	40	schlecht	**2,8**	,94	56	schlecht	**± 0**	,615 (,0)	**nein**
10	2	**1,4**	,58	50	gut	**1,8**	,60	36	gut	**- ,4**	,019 (,5)	**nein**
11	4	**2,2**	,63	47	mittel	**1,7**	,68	57	gut	**+ ,5**	**,000 (,6)**	**ja**
12	1	**2,0**	,63	40	gut	**2,3**	,70	32	mittel	**- ,3**	,028 (,4)	**nein**
13	6	**1,4**	,48	41	gut	**1,9**	,80	28	gut	**- ,5**	,009 (,6)	**nein**
14	1	**2,5**	,82	25	schlecht	**2,2**	,69	31	mittel	**+ ,3**	,120 (,4)	**ja**
15	3	**2,4**	,61	43	schlecht	**2,0**	,46	34	gut	**+ ,4**	**,000 (,5)**	**ja**
16	4	**1,7**	,66	29	gut	**1,9**	,67	39	gut	**- ,2**	,458 (,2)	**nein**
17	5	**2,7**	1,17	12	schlecht	**2,3**	,92	43	mittel	**+ ,4**	,895 (,5)	**ja**
18	3	**1,3**	,45	27	gut	**1,5**	,51	23	gut	**- ,2**	,275 (,2)	**nein**

Schulnoten: 1=sehr gut bis 6=ungenügend; * Das Signifikanzniveau wurde auf p<0,05 festgelegt. In Klammern befinden sich die Effektgrößen (d) der Mittelwertsunterschiede; # Angelegt wurde das Kriterium eines minimalen Mittelwertunterschieds in den Schulnoten von Δ=+0,255. In der Spalte „Signifikanz" sind die fünf in der zweiten Erhebungsphase signifikant verbesserten LSV fett markiert.
Liga: gut = Schulnote M ≤ 2,0; mittel = Schulnote M 2,1 ≤ M ≤ 2,3; schlecht = Schulnote M ≥ 2,4
(Acht Vorlesungen erfüllen das Kriterium eines minimalen Mittelwertunterschiedes in der Schulnote von Δ=+0,255. Dieser Unterschied ist jedoch nur in fünf Fällen signifikant.)

Tabelle 4: Gegenüberstellung der Auditoren-Bewertungen auf der Ebene der Einzelmerkmale aus den Erhebungsphasen 1 und 2 für die 18 erfassten Termine

	Erhebungsphase 1			Erhebungsphase 2			Signifikanz*
	M	SD	n	M	SD	n	
zur Vorlesung							
… Orientierung an Leitsymptomen	3,8	1,83	37	4,7	1,26	36	,063 (,6)
… Praxisbezug	4,8	0,89	37	5,1	0,83	36	**,048 (,3)**
… Anregung zum Mitdenken	3,9	1,61	37	4,8	1,18	36	**,018 (,6)**
… interdisziplinäre Bezüge	2,5	1,36	35	2,9	1,39	36	,292 (,3)
… leitsymptombezogenes Fachwissen	4,4	1,61	36	4,6	1,17	36	,981 (,2)
… strukturierter Aufbau	4,9	1,07	37	4,8	1,12	36	,659 (,1)
… Entsprechung des LSV-Konzepts	3,4	1,76	36	4,4	1,33	36	**,015 (,6)**
zur Lehrperson							
… vorbereitet	5,1	1,04	37	5,4	0,93	36	,188 (,3)
… freundlicher Umgang	5,0	1,12	36	5,1	1,33	36	,702 (,1)
… klarer Ausdruck	5,3	0,87	37	5,3	0,75	36	,853 (,0)
… akustisch verständlich	5,3	1,03	37	5,1	1,09	36	,283 (,2)
… lebendiger Vortrag	4,1	1,54	37	4,6	1,13	36	,372 (,4)
… interaktive Gestaltung	3,8	1,82	37	4,8	1,27	36	**,026 (,6)**
… Eingehen auf Zuhörer	4,6	1,54	34	4,9	1,17	36	,538 (,2)
… anschauliche Darstellung	4,2	1,51	37	5,0	1,16	36	**,015 (,6)**
… Bemühung um Lernerfolg	4,5	1,37	37	5,2	0,84	36	**,013 (,6)**
Schulnote	2,5	1,00	39	2,3	1,00	36	,288 (,2)

Likert-Skala: 1=trifft gar nicht zu bis 6=trifft sehr zu;
* Das Signifikanzniveau wurde auf p<0,05 festgelegt. In Klammern befinden sich die Effektgrößen (d) der Mittelwertsunterschiede. Fett markiert sind die in der zweiten Erhebungsphase signifikant besser bewerteten Merkmale, von denen je drei auf die Vorlesung selbst und auf die Lehrperson entfallen. Zwei wesentliche Aspekte sind hierbei die „Entsprechung des LSV-Konzepts" und die „interaktive Gestaltung".

Im Vergleich der Bewertungen der Einzelmerkmale durch die Studierenden und Auditoren (siehe Tabelle 5) ergeben sich für die erste Erhebungsphase für fast alle Merkmale durchweg statistisch signifikante Mittelwertunterschiede zwischen den beiden Gruppen, wobei die Auditoren die Vorlesungen insgesamt kritischer bewerten als die Studierenden. In der zweiten Erhebungsphase findet sich nur noch bei einem Merkmal ein signifikanter Unterschied zwischen den Bewertungen von Auditoren und Studierenden.

Übereinstimmung der Auditorenbewertungen

Wie die Berechnungen des Intraklassenkoeffizienten und der Signifikanzen ergaben, liegt die Übereinstimmung der Gruppe der Auditoren zwischen PJ-Studierenden und wissenschaftlichen Mitarbeitern auf Basis der Einzelmerkmale in der Erhebungsphase zwischen i_{CCmin}=-0,030 und i_{CCmax}=0,605 (siehe Tabelle 6). Beim überwiegenden Teil der erfassten Merkmale ist der Zusammenhang positiv signifikant. Im Gegensatz zu den Pilottestungen [32], bei denen v.a. bei den konzeptbezogenen Merkmalen große Differenzen zwischen den beiden Auditorengruppen bestanden, fällt die Übereinstimmung in der ersten Erhebungsphase sehr zufriedenstellend aus. In der zweiten Erhebungsphase liegt die Übereinstimmung der Auditoren zwischen i_{CCmin}=-0,022 und i_{CCmax}=0,771 und ist ebenfalls überwiegend positiv signifikant. Die Übereinstimmung zwischen den Auditorengruppen ist als mittelmäßig hoch zu bewerten, die Korrelationskoeffizienten weisen insgesamt eine recht breite Streuung auf.

Tabelle 6: Interklassenkorrelationskoeffizienten (mit Signifikanzen) für die Gruppe der Auditoren - Vergleich der PJ-Studierenden mit den wissenschaftlichen Mitarbeitern in den Erhebungsphasen 1 und 2

	PJ- und wissenschaftliche Auditoren	
	$i_{CCErhebungsphase1}$	$i_{CCErhebungsphase2}$
zur Vorlesung		
… Orientierung an Leitsymptomen	,504 (,000)	,709 (,000)
… Praxisbezug	,295 (,005)	**,245 (,151)**
… Anregung zum Mitdenken	,404 (,000)	,649 (,001)
… interdisziplinäre Bezüge	,379 (,001)	**,257 (,139)**
… leitsymptombezogenes Fachwissen	,433 (,000)	,383 (,049)
… strukturierter Aufbau	,217 (,029)	,655 (,001)
… Entsprechung des LSV-Konzepts	,602 (,000)	,771 (,000)
zur Lehrperson		
… vorbereitet	,394 (,000)	,751 (,000)
… freundlicher Umgang	**,047 (,344)**	**,021 (,463)**
… klarer Ausdruck	**-,030 (,602)**	,562 (,005)
… akustisch verständlich	,386 (,000)	**-,022 (,534)**
… lebendiger Vortrag	,484 (,000)	,487 (,015)
… interaktive Gestaltung	,605 (,000)	,592 (,003)
… Eingehen auf Zuhörer	,479 (,000)	,517 (,010)
… anschauliche Darstellung	,539 (,000)	,737 (,000)
… Bemühung um Lernerfolg	,374 (,000)	,497 (,014)
Schulnote	,548 (,000)	,620 (,002)

Fett markiert sind die Merkmale, bei denen sich keine signifikante Korrelation zwischen den PJ-Studierenden und den wissenschaftlichen Mitarbeitern ergab. Diese hat sich von zwei Merkmalen in der ersten Erhebungsphase auf vier in der zweiten Erhebungsphase erhöht.

Tabelle 5: Mittelwertvergleiche zwischen den Bewertungen der Studierenden und Auditoren nach Einzelmerkmalen über alle LSV-Termine, Ligen und Themenblöcke hinweg für die Erhebungsphasen 1 und 2

	Erhebungsphase 1							Erhebungsphase 2						
	Studierende			Auditoren			Signifikanz*	Studierende			Auditoren			Signifikanz*
	M	SD	n	M	SD	n		M	SD	n	M	SD	n	
zur Vorlesung														
... Orientierung an Leitsymptomen	4,8	1,30	3308	3,8	1,77	154	**,000 (,8)**	4,9	1,31	872	4,7	1,26	36	,373 (,2)
... Praxisbezug	4,8	1,11	3319	4,6	1,13	153	**,027 (,2)**	5,0	1,09	872	5,1	0,83	36	,593 (,1)
... Anregung zum Mitdenken	4,3	1,29	3333	3,9	1,52	155	**,003 (,3)**	4,7	1,28	872	4,8	1,18	36	,648 (,1)
... interdisziplinäre Bezüge	3,5	1,45	3283	2,6	1,52	148	**,001 (,6)**	3,6	1,57	872	2,9	1,39	36	**,013 (,5)**
... leitsymptombezog. Fachwissen	4,8	1,12	3297	4,3	1,60	151	**,002 (,4)**	4,8	1,26	872	4,6	1,17	36	,348 (,2)
... strukturierter Aufbau	5,0	1,03	3326	4,9	1,08	154	,242 (,1)	4,8	1,25	872	4,8	1,12	36	,993 (,0)
... Entsprechung des LSV-Konzepts	4,6	1,29	3244	3,4	1,77	152	**,000 (,9)**	4,5	1,66	872	4,4	1,33	36	,723 (,1)
zur Lehrperson														
... vorbereitet	5,4	0,87	3338	5,1	1,01	155	**,014 (,2)**	5,4	0,99	872	5,4	0,93	36	,991 (,0)
... freundlicher Umgang	5,3	0,92	3329	5,1	0,93	152	,068 (,2)	5,4	0,96	872	5,1	1,33	36	,069 (,3)
... klarer Ausdruck	5,3	0,89	3335	5,1	0,94	155	**,011 (,2)**	5,2	1,06	872	5,3	0,75	36	,553 (,1)
... akustisch verständlich	5,3	1,01	3337	5,0	1,14	155	**,004 (,3)**	4,9	1,39	872	5,1	1,09	36	,374 (,2)
... lebendiger Vortrag	4,5	1,25	3330	4,1	1,40	155	**,004 (,3)**	4,7	1,32	872	4,6	1,13	36	,546 (,1)
... interaktive Gestaltung	4,1	1,49	3328	3,7	1,71	155	**,003 (,3)**	4,6	1,36	872	4,8	1,27	36	,389 (,2)
... Eingehen auf Zuhörer	4,8	1,11	3173	4,5	1,59	142	**,003 (,3)**	4,8	1,43	872	4,9	1,17	36	,674 (,1)
... anschauliche Darstellung	4,6	1,24	3331	4,2	1,51	155	**,003 (,3)**	4,9	1,36	872	5,0	1,16	36	,662 (,1)
... Bemühung um Lernerfolg	4,8	1,04	3310	4,6	1,22	153	**,022 (,2)**	4,9	1,21	872	5,2	0,84	36	,081 (,3)
Schulnote	2,2	0,85	3247	2,6	0,97	155	**,003 (,5)**	2,1	0,81	872	2,3	1,00	36	,154 (,2)

Likert-Skala: 1=trifft gar nicht zu bis 6=trifft sehr zu; * Das Signifikanzniveau wurde auf p<0,05 festgelegt. In Klammern befinden sich die Effektgrößen (d) der Mittelwertsunterschiede. Fett markiert sind die Merkmale, in denen die Bewertungen der Studierenden und der Auditoren signifikant differieren. In der ersten Erhebungsphase trifft dies fast für alle Merkmale zu. In der zweiten Erhebungsphase ist dies nur noch bei einem Merkmal zutreffend.

Diskussion

Die Ergebnisse der Audits und Terminevaluationen in der ersten Erhebungsphase zeichnen ein insgesamt positiveres Bild der LSV als die im Vorfeld erhobenen studentischen Beurteilungen in der Trimesterabschlussevaluation hatten erwarten lassen. Hierbei könnte es sich um eine tatsächliche Verbesserung handeln. Es ist jedoch zu berücksichtigen, dass retrospektive, zusammenfassende Evaluationen tendenziell schlechter ausfallen als Evaluationen, die direkt im Anschluss an eine Veranstaltung erhoben werden [31], so dass diese Beobachtung auch durch einen methodischen Effekt erklärt werden könnte. Auf Basis der Schulnoten konnte die hypothetische Verbesserung der LSV-Gesamtbewertung in der zweiten Erhebungsphase nach der Intervention nur in mäßigem Ausmaß festgestellt werden. Das geforderte Verbesserungskriterium wurde in der Terminevaluation nur bei 44% der LSV erreicht. Bei den Auditoren fanden sich sogar nur in 28% der Evaluationen verbesserte LSV. Diesem Ergebnis steht jedoch die Bewertung der Auditoren auf Ebene der Einzelmerkmale gegenüber, die ganz überwiegend positivere Beurteilungen, vor allem der didaktischen Eigenschaften der Lehrpersonen in der zweiten Erhebungsphase zeigt. Eine Schwäche liegt hier in der geringen Gesamtfallzahl, die durch die initiale Ziehung der Stichproben zu einem Teil ausgeglichen wird.

Bei dem gewählten Veränderungskriterium der zu vergebenden Schulnote handelt es sich um ein relativ abstraktes Maß. Es lässt sich daher vermuten, dass diese Größe zu wenig differenziert ist, um eventuell bestehende Unterschiede der LSV nach der Intervention abzubilden, da es sich bei dem Konstrukt „Lehrqualität“ um ein komplexes Merkmal handelt [14]. Für den Verlust von Information durch den Einsatz von Schulnoten spricht außerdem die Diskrepanz bei der Gruppe der Auditoren zwischen der summativen Kenngröße der Note und den parallel bewerteten Einzelmerkmalen, die deutliche Verbesserungen zeigen. In der ersten Erhebungsphase bewerten die geschulten Auditoren die LSV in fast allen Merkmalen signifikant kritischer als die teilnehmenden Studierenden, wie in Hypothese 2 vermutet. Die Bewertungen der Auditoren nach Einzelmerkmalen fallen in der zweiten Erhebungsphase im Gegensatz zu den studentischen Erhebungen wesentlich besser aus. Es könnte sich dabei einerseits um eine tatsächliche qualitativ-didaktische Verbesserung der LSV handeln, die von geschulten Auditoren differenzierter wahrgenommen und bewertet wurde. Andererseits muss auch ein möglicher Einfluss des Rosenthal-Effekts berücksichtigt werden [24], wodurch die bloße Erwartung einer Verbesserung der LSV nach der Intervention bei den Auditoren zu einer besseren Bewertung geführt haben könnte. Jedoch wird der Einsatz von geschulten Auditoren als für eine valide und forschungspraktikable Beurteilung von Lehrqualität beschrieben [1], [17]. Auch in anderen Arbeiten finden sich teilweise nur moderate Übereinstimmungen von studentischen und „peer-Bewertungen“ [16]. In der zweiten Erhebungsphase fallen die Unterschiede in der Bewertung weniger deutlich aus, was insbesondere innerhalb der Gruppe der Auditoren für eine homogenere Bewertungsgrundlage nach der erfolgten Schulung sprechen könnte. Die so nachgewiesene hohe Inter-Rater-Reliabilität stützt die Validität der Daten [34].

Weiterhin muss analysiert werden, ob die für das Projekt gewählte Intervention zur Verbesserung der LSV stark genug war. Da die Literatur keine Evidenz dafür liefert, dass studentische Evaluation allein die Lehre an Hochschulen verbessert [23], [28], wurde in dieser Studie eine über bloßes Ergebnisfeedback hinausgehende Intervention gewählt. Das Feedback an die Zielgruppe erfolgte jedoch nur in schriftlicher Form. Andere Untersuchungen zeigen, dass schriftliche Rückmeldungen von Lehrenden nur selten gelesen werden und damit kaum Auswirkungen haben können [9]. Weitere Bemühungen wie z.B. die Durchführung von hochschuldidaktischen Beratungen [23], [33] oder direkte Diskussionen mit den Lehrenden

über die Ergebnisse [2] ziehen hingegen wirkungsvollere Verbesserungen nach sich. Außerdem erhöht ein möglichst früher Zeitpunkt des Feedbacks die Wahrscheinlichkeit eines positiven Effekts bei den Lehrenden [26]. In der vorliegenden Untersuchung war der Zeitraum zwischen Erhebung und Rückmeldung mit bis zu vier Monaten vergleichsweise lang. Allerdings wurden die personenbezogenen schriftlichen Rückmeldungen, wie im Methodenteil beschrieben, anschaulich aufbereitet und eingehend erläutert. Bekannt ist nämlich, dass schriftliche Rückmeldungen von Evaluationen ohne Erläuterungen häufig von Lehrenden nicht korrekt interpretiert und somit nicht richtig verstanden werden und daher meist wirkungslos bleiben [2]. Ein weiterer Einflussfaktor für den eher schwachen Effekt der Intervention könnte in der Tatsache begründet liegen, dass es sich bei der LSV um eine „multi-instructor-Veranstaltung“ mit insgesamt ca. 150 Lehrpersonen in sechs Themenblöcken handelt. Ein solches Format birgt bei der Umsetzung von Veränderungen oder Verbesserungen im Vergleich mit Kursen, bei denen lediglich wenige oder gar nur einzelne Personen beteiligt sind, besondere Schwierigkeiten [26]. Weiterhin ist bekannt, dass die in Evaluationsprojekten angebotenen Informationen und Beratungen von Lehrenden weniger genutzt werden, wenn diese nicht daran interessiert oder gewillt sind, ihre didaktischen Fertigkeiten zu verbessern [18].

Ein weiterer Aspekt für den nicht sehr durchgreifenden Effekt der Intervention ist in der systemimmanenten Trägheit von Fakultäten bei der Umsetzung von curricularen Innovationen zu vermuten [27]. Hinzu kommt, dass das Konzept für die LSV bis zur Intervention im Rahmen dieser Studie nicht während der curricularen Planungen schriftlich fixiert und an die Lehrenden übermittelt wurde. Damit wurde der Faktor „Kommunikation innerhalb der Fakultät“, der in Planungsprozessen von wesentlicher Bedeutung ist [3], bei der Einführung des neuen Curriculums nicht ausreichend beachtet. Idealer wäre es, eine Schulungsmaßnahme anzustreben, die alle an der LSV beteiligten Lehrpersonen mit dem Konzept vertraut macht [3]. In die anschließenden Maßnahmen zur Überprüfung der Qualität der LSV sind, wie in dieser Studie erfolgt, Lehrende und Studierende einzubinden, um in der Fakultät eine möglichst hohe Akzeptanz zu erzielen [19].

Um insgesamt die Effektivität von Lehr- und Lehrveranstaltungsevaluationen zu steigern, sind diese in ein allgemeines Verfahren zur Bestimmung und Förderung von Lehr-, Ausbildungs- und Forschungsqualität zu integrieren, da eine alleinige Einschätzung von Lehrqualität für eine Verbesserung nicht ausreichend ist [25].

Zusammenfassung und Ausblick

Die vorliegende Untersuchung konnte zeigen, dass sich in der Evaluation eines neu etablierten Konzepts der LSV sowohl in der Terminevaluation durch Studierende als auch durch geschulte Auditoren nach einer Intervention didaktische Verbesserungen nachweisen ließen, die sich stärker auf der Basis differenzierter Einzelmerkmale zu lehrpersonen- und konzeptbezogenen Merkmalen als durch Schulnoten abbilden ließen. Die Studierenden bewerteten die LSV insgesamt positiver als die Auditoren, wobei eine gute Inter-Rater-Reliabilität bestand. Eine dreistündige Vorbereitung der Auditoren reicht offenbar jedoch nicht aus, um die Personen adäquat auf ihre Rolle als analysierende Feedbackgeber vorzubereiten. Außerdem ist zu berücksichtigen, dass die Generalisierbarkeit der Ergebnisse aufgrund der methodisch bedingten Stichprobenwahl mit nur 18 Vorlesungen in der zweiten Erhebungsphase eingeschränkt ist. Die Notwendigkeit der besseren inhaltlichen und strukturellen Einbettung der LSV in das curriculare Gesamtkonzept auch im Sinne einer Begleitung durch eine regelmäßige Qualitätskontrolle wurde in dieser Studie dennoch deutlich. Wie lange Wirkungen, die aufgrund des Feedbacks nach einer Evaluation eintreten, bei der Zielgruppe bestehen bleiben, sollten zukünftige Studien untersuchen. Eine rein schriftliche Information zum Design von Vorlesungen nach modernen didaktischen Kriterien scheint als Intervention für viele Dozierende kein ausreichender Stimulus zur Verbesserung oder Überarbeitung ihrer Vorlesungen zu sein. Weiterhin ist auch zu prüfen, welche Effekte bei der indirekt betroffenen Gruppe der Studierenden eintreten, z.B. Auswirkungen auf ihre Motivation und ihren Lernerfolg.

Danksagung

Wir danken der Medizinischen Fakultät der Universität Hamburg für die Förderung dieses Projekts (L-107/2006) aus dem Förderfonds Lehre.

Interessenkonflikt

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

Literatur

1. Albanese MA, Schuldt SS, Case D, Brown D. The validity of lecturer ratings by students and trained observers. Acad Med. 1991;66(5):26-28. DOI: 10.1097/00001888-199101000-00008
2. Baggott J. Reaction of lecturers to analysis results of student ratings of their lecture skills. J Med Educ. 1987;62:491-496.
3. Bland CJ, Starnaman S, Wersal L, Moorhead-Rosenberg L, Zonia S, Henry R. Curricular change in medical schools: how to succeed. Acad Med. 2000;75(6):575-594. DOI: 10.1097/00001888-200006000-00006
4. Bortz J, Döring N. Forschungsmethoden und Evaluation. Berlin: Springer; 2006.
5. Brown G, Manogue M. AMEE Medical Education Guide No. 22: Refreshing lecturing: a guide for lecturers. Med Teach. 2001;23(3):231-244. DOI: 10.1080/01421590120043000

6. Butler JA. Use of teaching methods within the lecture format. Med Teach. 1992;14(1):11-23. DOI: 10.3109/01421599209044010
7. Cantillon P. Teaching large groups. BMJ. 2003;326:437-440.
8. Clauß G, Ebner H. Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Thun/Frankfurt a. M.: Harri Deutsch; 1977.
9. Cohen PA. Effectiveness of student-rating feedback for improving college instruction: a meta-analysis of findings. Res High Educ. 1980;13(4):321-341. DOI: 10.1007/BF00976252
10. Copeland H, Longworth D, Hewson M, Stoller J. Successful lecturing. A prospective study to validate attributes of the effective medical lecture. J Gen Intern Med. 2000;15(6):366–371. DOI: 10.1046/j.1525-1497.2000.06439.x
11. Craig M. Facilitated student discussions for evaluating teaching. SIGCSE Bulletin. 2007;39(1):190-194. DOI: 10.1145/1227504.1227376
12. Diehl JM. Normierung zweier Fragebögen zur studentischen Beurteilung von Vorlesungen und Seminaren. Psychol Erz Unterr. 2003;50:27-42.
13. Fyrenius A, Bergdahl B, Silén C. Lectures in problem-based learning - why, when and how? An example of interactive lecturing that stimulates meaningful learning. Med Teach. 2005;27(1):61-65. DOI: 10.1080/01421590400016365
14. Gordon PA. Student evaluation of college instructors: an overview. Valdosta: Valdosta State University; 1997. Zugänglich unter/available under: http://teach.valdosta.edu/WHuitt/files/tcheval.pdf
15. Grass G, Stosch C, Griebenow R. Renaissance der Vorlesung. Dtsch Ärztebl. 2005;102(23):A1642.
16. Greenwood GE, Ramagli HJ. Alternatives to student ratings of college teaching. J High Educ. 1980;51(6):673-684. DOI: 10.2307/1981172
17. Imseis HM, Galvin SL. Faculty and resident preference for two different forms of lecture evaluation. Am J Obstet Gynecol. 2004;191(5):1815-1821. DOI: 10.1016/j.ajog.2004.07.068
18. Irby D, DeMers J, Scher M, Matthews D.A model for the improvement of medical faculty lecturing. J Med Educ. 1976;51(5):403-409.
19. Leppek R, Jußen M, Berthold D, Sulzer J, Klose KJ. Windmühlenprinzip versus Uhrwerkprinzip - Tradition und Interaktion in der akademischen Vorlesung. Z Ärztl Fortbild. 1996;90:406-413.
20. Moßig I. Stichproben, Stichprobenauswahlverfahren und Berechnung des minimal erforderlichen Stichprobenumfangs. Gießen: Universität Gießen;1996.
21. Reed M. Electronic module evaluation: combining quality with quantity. Kongressbeitrag University of Leeds Inaugural Learning and Teaching Conference. Leeds: University of Leeds; 2004. Zugänglich unter/available under: http://homepages.see.leeds.ac.uk/~lecmsr/Reed%202004.doc
22. Rindermann H. Methodik und Anwendung der Lehrveranstaltungsevaluation für die Qualitätsentwicklung an Hochschulen. Sozialwis Berufspraxis. 2003;26(4):401-413.
23. Rindermann H. Quality of instruction improved by evaluation and consultation of instructors. Int J for Acad Develop. 2007;12(2):73-85. DOI: 10.1080/13601440701604849
24. Rost DH. Handwörterbuch der Pädagogischen Psychologie. Weinheim: Beltz; 2001.
25. Schmidt B. Warum oft wirksam? Und warum manchmal wirkungslos? – Subjektive Erklärungen zur Wirkung von Lehrveranstaltungsevaluation aus der Sicht von Nutzern und Anbietern. Z Eval. 2008;7(1):7-33.
26. Stillman PL, Gillers MA, Heins M, Nicholson G, Sabers D. Effect of immediate student evaluations on a multi-instructor course. J Med Educ. 1983;58:172-178.
27. Sukkar MY. Curriculum development: a strategy for change. Med Educ. 1986;20:301-306. DOI: 10.1111/j.1365-2923.1986.tb01369.x
28. Turhan K, Yaris F, Nural E. Does instructor evaluation by students using a web-based questionnaire impact instructor performance? Adv Health Sci Educ. 2005;10(1):5-13. DOI: 10.1007/s10459-004-0943-7
29. Universität Hamburg. Hamburger Lernzielkatalog. Hamburg: Universität Hamburg; 2009. Zugänglich unter/available under: http://www.uke.de/studierende/downloads/zg-studierende/Lernzielkatalog_091104_mat.pdf
30. van den Bussche H, Anders S, Ehrhardt M, Göttsche T, Hüneke B, Kohlschütter A, Kothe R, Kuhnigk O, Neuber K, Rijntjes M, Quellmann C, Harendza S. Lohnt sich eine Reform der klinischen Ausbildung? - Die Qualität des Hamburger Curriculums unter der alten und der neuen Approbationsordnung im Vergleich. Z Ärztl Fortbild Qualitätssich. 2005;99:419-423.
31. van den Bussche H, Weidtmann K, Kohler N, Frost M, Kaduskiewicz H. Evaluation der ärztlichen Ausbildung: Methodische Probleme der Durchführung und der Interpretation von Ergebnissen. GMS Z Med Ausbild. 2006;23(2):Doc37. Zugänglich unter/available under: http://www.egms.de/de/journals/zma/2006-23/zma000256.shtml
32. Weidtmann K. Analyse des Status quo der Leitsymptom-Vorlesung und Planung einer evaluationsbasierten Intervention an der Medizinischen Fakultät Hamburg. Unveröffentlichte Projektarbeit im Studiengang Master of Medical Education. Heidelberg: Medizinische Fakultät Heidelberg; 2007.
33. Wilson RC. Improving faculty teaching: Effective use of student evaluations and consultants. J High Educ. 1986;57(2):196-211. DOI: 10.2307/1981481
34. Wirtz M. Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen. Rehabilitation. 2004;43:384-389. DOI: 10.1055/s-2003-814935

Korrespondenzadresse:
Dr. med. Olaf Kuhnigk, MME (Bern)
Universitätsklinikum Hamburg-Eppendorf, Klinik für Psychiatrie und Psychotherapie, Martinistraße 52, 20246 Hamburg, Deutschland, Tel.: +49 (0)40/7410-57675, Fax: +49 (0)40/7410-54702
o.kuhnigk@uke.de

Bitte zitieren als
Kuhnigk O, Weidtmann K, Anders S, Hüneke B, Santer R, Harendza S. Leitsymptomvorlesungen im klinischen Studienabschnitt - Effekte evaluationsbasierter Interventionen auf eine Großgruppen-Lehrveranstaltung. GMS Z Med Ausbild. 2011;28(1):Doc15.
DOI: 10.3205/zma000727, URN: urn:nbn:de:0183-zma0007272

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2011-28/zma000727.shtml

Eingereicht: *15.01.2010*
Überarbeitet: *02.08.2010*
Angenommen: *23.09.2010*
Veröffentlicht: *04.01.2011*

Lectures based on cardinal symptoms in undergraduate medicine - effects of evaluation-based interventions on teaching large groups

Abstract

Despite critical voices lectures are still an important teaching format in current medical curricula. With the curricular reform at Hamburg Medical Faculty in the year 2004, all subject specific lectures were replaced by cardinal symptom oriented lectures (LSV) in the new clinical curriculum. LSVs are taught throughout all six thematic blocks in years three to five. Since regular student evaluations after each thematic block seemed to demand improvement of the LSVs, this study was carried out using evaluations of individual LSVs by the participating students and by trained auditors (final year students and academic staff). Based on these evaluations feedback containing the individual evaluation data was given in written form to the lecturers combined with information material on planning an LSV using modern didactic techniques. In a second evaluation period, the effects of this intervention were studied. Only small improvements in the LSVs' quality were noted regarding the level of marks achieved. When individual items were evaluated, especially the didactic quality, significant improvements were noticeable. Overall, on the basis of individual items students ranked the quality of the LSVs significantly higher than trained auditors during the first evaluation period. This effect was no longer seen after the second evaluation period. The inter rater reliability among the auditors was very good. This study shows that regular quality assurance is needed on the structural levels and for staff to accompany the process of embedding teaching formats into curricular concepts. Further investigation is needed to determine the adequate frequency of evaluation and the format of feedback to guarantee sustainable effects of the didactic quality of lectures.

Keywords: Lecture, evaluation, audit, intervention, guideline, didactic skills, faculty development, quality assurance

Olaf Kuhnigk[1,2]
Katja Weidtmann[2]
Sven Anders[3]
Bernd Hüneke[4]
René Santer[5]
Sigrid Harendza[6]

1 Universitätsklinikum Hamburg-Eppendorf, Klinik für Psychiatrie und Psychotherapie, Hamburg, Deutschland

2 Universitätsklinikum Hamburg-Eppendorf, Prodekanat für Lehre, Hamburg, Deutschland

3 Universitätsklinikum Hamburg-Eppendorf, Institut für Rechtsmedizin, Hamburg, Deutschland

4 Universitätsklinikum Hamburg-Eppendorf, Klinik und Poliklinik für Geburtshilfe und Pränatalmedizin, Hamburg, Deutschland

5 Universitätsklinikum Hamburg-Eppendorf, Klinik und Poliklinik für Kinder- und Jugendmedizin, Hamburg, Deutschland

6 Universitätsklinikum Hamburg-Eppendorf, III. Medizinische Klinik, Hamburg, Deutschland

Introduction

Lectures as a learning format

The teaching format „lecture" is, despite criticism and subsequent curricular reforms, still an important didactic element in undergraduate medical education [6]. On the one hand, traditional systematic lectures are criticised for not promoting the development of independent thinking. On the other hand, they provide an opportunity to impart information to groups of learners in an economical and resource-efficient way, to deliver an introduction to complex topics, and to describe current research results and personal, clinical or scientific experiences [5].

To benefit form the potential advantages of lectures as a teaching format they should be blended into a curricular framework [15] and linked with other teaching formats to stimulate students' learning on their own accord [13]. In this respect, a case based format has proved its value [10].

Cardinal symptom-based lectures in the Hamburg curriculum

A comprehensive reform of the clinical part of the undergraduate medical curriculum at the medical faculty of the University of Hamburg was carried out in 2004 with the main focus of the reformed clinical curriculum in medicine (KliniCuM) being on better integration of the subjects and greater practical educational components [30]. Lessons of curricular years three to five are distributed in six thematic blocks and one elective block and content is organized according to the Hamburg catalogue of learning objectives [29] which depicts the different dimensions of learning and levels of competencies. Systematic and subject-specific lectures were abolished during the reform process and replaced by lectures which are case-based, with their contents geared to the cardinal symptoms of different diseases. The concept of these cardinal symptom-oriented lectures (LSV) is an integral part of KliniCuM which runs through all thematic blocks as a thread and reveals the content links to the other learning sessions (problem based tutorials, bedside teaching). First measurements of participant numbers and evaluation data promise greater student satisfaction with this new lecture format as compared with the lecture format prior to the reform. A need for further improvement was still visible shortly after the reform, yet concrete points of critique were formulated by the students in mostly global commentaries in the evaluations held at the end of every thematic block and remained somewhat unclear [32].

Evaluation as basis for intervention

Students have the ability to evaluate the didactic quality of courses in a valid and reliable way [12], [22]. At the same time there is a high demand to base the evaluation of teaching not only on student judgement [11], [21]. Since the LSV was designed to use the value of the teaching format "lecture" according to the above mentioned criteria [10], [13], we performed this study for quality control. This included a detailed investigation of the critiques and shortcomings mentioned in the evaluations and the observations of the impact of intervention based on this analysis regarding the improvement of this teaching format.

Methods

Question and hypotheses

Two questions were investigated in our study. One: Are there noticeable differences in the evaluation of the LSV after an intervention based on previous evaluation results, when the new LSV is evaluated by participating students and trained auditors? Two: Is there a difference between the assessment values of participating students and trained auditors?
The main hypotheses are:

1. The LSV will receive more positive ratings after the intervention, especially regarding its didactic values.
2. The evaluation of trained auditors will be more consistent and all in all more critical than the evaluation of the participating students.

Instruments

Figure 1 shows an overview of the study and includes two evaluation phases and one intervention phase. A checklist for the audits of the LSV was designed and validated in a pilot phase [32]. This checklist includes seven items regarding structure and content of the lecture as well as nine items regarding the didactic qualities of the lecturers. The group of auditors comprises eight physicians and scientists and 14 students in the final year of their undergraduate studies. Pairs of auditors (one physician and one student) were allotted to control for possible systemic differences, e.g. diverging perspectives because of status (non-student/student). Prior to the pilot phase, all auditors underwent a three hour training session, including explanation of the instruments and rehearsal of standardised methodical means for the evaluation.
In addition, a questionnaire was developed for the students who participated in the audited LSV. This questionnaire included central aspects of the LSV such as focus on cardinal symptoms, practical applications, and structural design of the lecture as well as items regarding the lecturer, e.g. manner of contact with the students, comprehensibility and clarity of the lecture. Furthermore, characteristics of the students, e.g. gender and continuity of visiting the LSV were also documented. The items in the questionnaires and checklists were rated on a 6-point Likert-scale (1: "I strongly agree" to 6: "I strongly disagree"). In addition, free commentaries were also possible and an overall rating of the lecture following the school grade system was given (1=very good, 2=good, 3=satisfactory, 4=sufficient, 5=poor, 6=deficient). In the pilot, the instruments were found to be feasible and comprehensible. The only minor modifications required for the final study such as readjusting the question sequence and clarification of some items by giving an example in brackets.
Since the pilot phase revealed considerable differences between the two groups of auditors (physicians/students), it was assumed that the initial training of the auditors

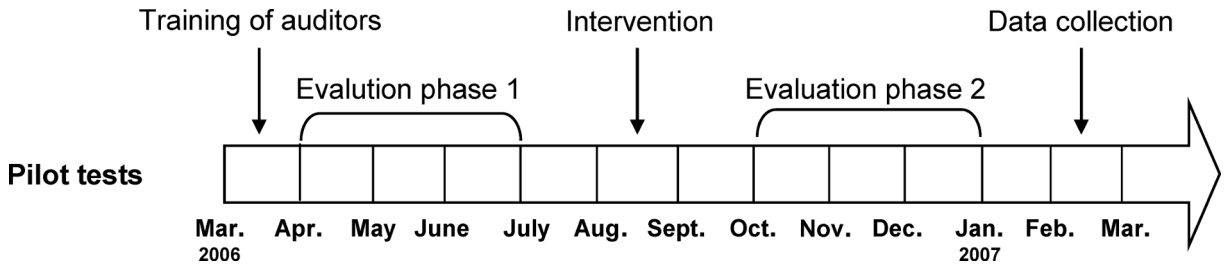

Figure 1: Timeline of the study with an overview of the individual steps

had not covered all main aspects of the study sufficiently. Therefore, a second training session took place before the first evaluation phase, including a summary of the LSV concept within KliniCuM, a delineation of the orientation along cardinal symptoms and connections between the subjects of a thematic block as well as cardinal symptom-oriented teaching of expert knowledge based upon concrete examples. The concordance of the ratings between the two groups of auditors was determined using the intra-class correlation coefficient (ICC) [34].

Design and sampling

All lecturers participating in the LSV were informed about this study. They were not notified though whether or which of their lectures had been randomly selected for an evaluation. For the first evaluation phase, a randomised representative sample of about one third of the LSVs per thematic block (altogether n=85 from all thematic blocks) was chosen from a total of 247 individual LSVs from all six thematic blocks during the trimester April to July 2006. This approach is classified as "drawing of stratified lots" and was chosen because the population parameters (all LSVs from all thematic blocks) were considered to be very heterogeneous, meaning that the feature characteristics of the basic population could exhibit major differences.
To reproduce all shades of the basic population sufficiently in a random sample, this sample needs to be very large according to the principle of a mere random selection to ensure representativeness. To address this problem, the basic population was divided into disjunctive classes (layers). It was assumed that the elements of each class will behave similarly regarding the research question and elements from different classes would be defined by their different characteristics [8]. In our study the classes are defined by the six thematic blocks. From proportionally layered samples, a random sample was drawn from each class by drawing individual lectures. SPSS 16.0 was used for the statistical evaluation of the data. Means were calculated by t-test for independent samples (level of significance $p<0.05$). The statistical tests were used descriptively.
Based on the data from the first evaluation phase the following intervention took place. Three groups were identified to receive an intervention: the teachers who were evaluated in the random sample (group 1), all teachers and all heads of departments participating in the LSV who had not yet been evaluated (group 2), students and the general public (group 3). Components of the intervention were:

- Letters: all members of groups 1 and 2 were sent a personal letter describing background, approach and goals of the project as well as the respective feedback components and a contact person for questions.
- General feedback: all three groups received the analysis of the data of the first evaluation phase with the non-personal, general statistics.
- Individual feedback: all members of group 1 received their personal feedback with the rating and the free commentaries of auditors and students.
- LSV manual: all groups received a manual based on the data of the first evaluation phase as a gold standard for the design of an LSV including concrete hints regarding content and form.
- Publication: groups 1 and 2 received the paper "Teaching large groups" [7], a publication about the design of teaching formats for large groups in medical education which contains easy-to-accomplish suggestions for the design of lectures following modern didactic concepts in a very consolidated description.
- For group 3 information regarding the LSV and the LSV manual were displayed on the homepage of the dean of education's office in the internet.

All evaluated lectures form the first evaluation phase (n=78, the smaller number is explained by three cancelled audits and four missing lectures) were subdivided in three leagues on the basis of their marks according to German school grades for the second evaluation phase (see table 1). Selection of the LSVs which were to be evaluated was realized by drawing quota-samples [4], meaning a deliberate selection which aspires for the sample to be drawn to simulate conditions of the basic population. In our case the basic population is defined by the six thematic blocks and the three leagues in which the lectures were selected representatively. Since our study particularly focused on the question whether the didactic quality of the LSV improved with the above mentioned intervention, this was operationalized by certain criteria like the students' evaluation of the LSV by school grades at the respective points in time. Considerations regarding the smallest difference desired in the students' ratings as well as the power analysis a required sample size of n=633 was identified [20]. In this case the following relevant criteria were achieved: effect size d=0.3 (d=0.1: small effect, d=0.3: medium effect, d=0.5 large effect), minimal difference of the means $\Delta=+0{,}255$, test power $1-\beta=0.8$ and α

α=0.05. Since on average n=35 student ratings were collected during the first evaluation phase per LSV, n=18 lectures were required for the second evaluation phase. Fourteen lectures of the second evaluation phase were held by the same teachers as in the first evaluation phase. The topics of all 18 LSVs were exactly the same as in evaluation phase 1. To guarantee an even distribution of the measurements on the structure of the basic population one lecture from each of the six thematic blocks and every league was chosen.

Table 1: Distribution of the results of the LSV of evaluation phase 1 on the three leagues for the selection of lectures for evaluation phase 2

League	Rating	$M_{School\ grade}$	LSV TB 1	LSV TB 2	LSV TB 3	LSV TB 4	LSV TB 5	LSV TB 6	LSV_{total}	$n_{Stud.}$
1	good	M ≤ 2,0	3	6	5	5	2	4	25	1017
2	medium	2,1 ≤ M ≤ 2,3	6	1	4	3	5	4	23	1161
3	bad	M ≥ 2,4	4	7	6	5	7	1	30	1176
LSV_{total}			13	14	15	13	14	9	78	
$n_{Stud.}$			630	634	726	595	348	421		3354

M: mean, LSV: cardinal symptom-oriented lecture, TB: thematic block, n: number
For the second evaluation phase the evaluated lectures (n=78) from the first evaluation phase were distributed in three leagues according to their ratings in German school grades.

Results

Changes in the rating of LSV as per school grades

On the basis of school grades comparisons of the means of the student ratings of the same 18 LSV rated in evaluation phases 1 and 2 five show significant improvements (28%), three significant deteriorations (17%), and ten ratings are found unchanged (55%) (see table 2). Hence, the majority of student evaluations of the LSVs in the second evaluation phase do not reveal changes. The above mentioned criterion of the minimal difference of the means in school grade of Δ=+0,255 is achieved by eight lectures (44%).

The evaluation of the LSV by school grade performed by the auditors does also not reveal an even picture regarding the effect of the intervention (see table 3). The percentage of the improvements matches the above mentioned student ratings of the individual lectures with four of five same LSVs rated as improved. In the first evaluation phase the physician and scientist auditors rated six lectures of 18 by one school grade lower than the student auditors, in the second evaluation phase seven.

Rating of didactics on the basis of individual items

A more differentiated picture compared to the one drawn by the school grades is shown by the comparison of individual items by the auditors from both evaluation phases (see table 4). The second evaluation phase reveals six significant improvements in ratings after the intervention and all other items except for three show a positive trend. In total, the improvements regarding the items "orientation to cardinal symptoms", "encouragement to follow the general train of thought", "use of LSV concept", "interactive design", "depictive presentation", and "effort to support successful learning" display large effect sizes. In the first evaluation phase the comparison between student and auditor ratings (see table 5) on the basis of individual items shows statistically significant differences between both groups for almost all items with the auditors rating the lectures more critically than the students. Ratings from the second evaluation phase reveal a significant difference between student and auditor ratings for only one item.

Table 3: Comparison of school grades given by the auditors for the 18 lectures from evaluation phases 1 and 2

No.	TB	Auditor-$SG_{Evaluation1}$		Auditor-$SG_{Evaluation2}$		Comparison#
		Stud	Phys	Stud	Phys	
1	1	1	2	2	3	worsened
2	5	3	4	3	4	unchanged
3	2	2	3	2	3	unchanged
4	3	2	2	4	4	worsened
5	5	3	3	2	3	unchanged
6	6	4	5	2	1	**improved**
7	2	2	2	1	1	**improved**
8	6	2	2	2	2	unchanged
9	4	4	4	2	3	**improved**
10	2	2	2	2	1	unchanged
11	4	4	3	4	2	unchanged
12	1	3	3	2	3	unchanged
13	6	2	1	1	1	unchanged
14	1	3	4	2	2	**improved**
15	3	3	2	2	1	**improved**
16	4	1	2	1	2	unchanged
17	5	3	3	4	4	worsened
18	3	2	2	2	2	unchanged

TB: Thematic block; SG: school grade, 1=very good to 6=deficient; Stud: student auditors, Phys: physician auditors.
improved: during the second evaluation phase school grades of both auditors were at least one school grade better of the school grade of one assessor was at least two grades better than during the first evaluation phase.
Worsened: During the second evaluation phase school grades of both auditors were at least one grade worse or the school grade of one assessor was at least two grades worse than during the first evaluation phase.
In the column "Comparison" the LSV which improved during the second evaluation phase are marked in bold. This is only the case for five LSV.

Table 4: Comparison of the auditor evaluations on the basis of individual items from the evaluation phases 1 and 2 for the 18 included lectures

	Evaluation phase 1			Evaluation phase 2			Significance*
	M	SD	n	M	SD	n	
Lecture							
... cardinal symptom-oriented	3,8	1,83	37	4,7	1,26	36	,063 (,6)
... practice oriented	4,8	0,89	37	5,1	0,83	36	**,048 (,3)**
... stimulates own thinking	3,9	1,61	37	4,8	1,18	36	**,018 (,6)**
... interdisciplinary connections	2,5	1,36	35	2,9	1,39	36	,292 (,3)
... symptom oriented knowledge	4,4	1,61	36	4,6	1,17	36	,981 (,2)
... good structure	4,9	1,07	37	4,8	1,12	36	,659 (,1)
... accordance with LSV-concept	3,4	1,76	36	4,4	1,33	36	**,015 (,6)**
Teacher							
... prepared	5,1	1,04	37	5,4	0,93	36	,188 (,3)
... friendly	5,0	1,12	36	5,1	1,33	36	,702 (,1)
... clear wording	5,3	0,87	37	5,3	0,75	36	,853 (,0)
... acoustically understandable	5,3	1,03	37	5,1	1,09	36	,283 (,2)
... lively talk	4,1	1,54	37	4,6	1,13	36	,372 (,4)
... interactive organisation	3,8	1,82	37	4,8	1,27	36	**,026 (,6)**
... consideration of participants	4,6	1,54	34	4,9	1,17	36	,538 (,2)
... graphic presentation	4,2	1,51	37	5,0	1,16	36	**,015 (,6)**
... effort for learning success	4,5	1,37	37	5,2	0,84	36	**,013 (,6)**
School grade	2,5	1,00	39	2,3	1,00	36	,288 (,2)

Likert-scale: 1=strongly disagree to 6=strongly agree;
* Significant p-value: p<0,05. Brackets show effect sizes (d) of differences to the mean. Bold are the items which were rated significantly better in the second evaluation phase, three regarding the lecture and three regarding the teacher. Two main aspects are "accordance with LSV-concept" and "interactive organisation".

Conformity of auditor ratings

As the calculation of the intra-class correlation coefficient and of the significances show the conformity of the ratings on the basis of individual items within the auditors between group of students in the final year and the group

Table 2: Comparison of the school grades given by the students in evaluation phase 1 und 2

No.	TB	Evaluation 1			League$_{Eva1}$	Evaluation 2			League$_{Eva2}$	$\Delta_{Evalutation\ 1\text{-}2}$	Significance*	Fulfillment of criterium#
		M_{SG}	SD	$n_{Stud.}$		M_{SG}	SD	$n_{Stud.}$				
1	1	**2,2**	,75	67	medium	**2,7**	,96	47	bad	**- ,5**	,001 (,6)	**no**
2	5	**2,3**	,53	56	medium	**2,4**	,87	56	bad	**- ,1**	,489 (,1)	**no**
3	2	**2,4**	,86	70	bad	**2,0**	,76	50	good	**+ ,4**	**,002 (,5)**	**yes**
4	3	**2,3**	,85	56	medium	**2,2**	,79	49	medium	**+ ,1**	,466 (,1)	**no**
5	5	**2,0**	,59	22	good	**2,1**	,72	54	medium	**- ,1**	,442 (,1)	**no**
6	6	**2,6**	,79	64	bad	**2,2**	,68	66	medium	**+ ,4**	**,008 (,5)**	**yes**
7	2	**2,1**	,62	67	medium	**1,9**	,61	78	good	**+ ,2**	**,004 (,2)**	**yes**
8	6	**2,2**	1,06	56	medium	**1,9**	,80	60	good	**+ ,3**	,076 (,4)	**yes**
9	4	**2,8**	,84	40	bad	**2,8**	,94	56	bad	**± 0**	,615 (,0)	**no**
10	2	**1,4**	,58	50	good	**1,8**	,60	36	good	**- ,4**	,019 (,5)	**no**
11	4	**2,2**	,63	47	medium	**1,7**	,68	57	good	**+ ,5**	**,000 (,6)**	**yes**
12	1	**2,0**	,63	40	good	**2,3**	,70	32	medium	**- ,3**	,028 (,4)	**no**
13	6	**1,4**	,48	41	good	**1,9**	,80	28	good	**- ,5**	,009 (,6)	**no**
14	1	**2,5**	,82	25	bad	**2,2**	,69	31	medium	**+ ,3**	,120 (,4)	**yes**
15	3	**2,4**	,61	43	bad	**2,0**	,46	34	good	**+ ,4**	**,000 (,5)**	**yes**
16	4	**1,7**	,66	29	good	**1,9**	,67	39	good	**- ,2**	,458 (,2)	**no**
17	5	**2,7**	1,17	12	bad	**2,3**	,92	43	medium	**+ ,4**	,895 (,5)	**Yes**
18	3	**1,3**	,45	27	good	**1,5**	,51	23	good	**- ,2**	,275 (,2)	**no**

School grades: 1=very good to 6=deficient; * significance was set to p<0,05. Brackets show effect sizes (d) as differences of the mean; # The criterion of a minimal differences in the means of school grades was set to Δ=+0,255. The column "Significance" shows the significantly improved LSVs in bold. Eva: evaluation, SG: School grade, TB: thematic block, n: number, League: good = school grade M ≤ 2,0; medium = school grade M 2,1 ≤ M ≤ 2,3; bad = school grade M ≥ 2,4. Eight lectures fulfilled the criterion of a minimal difference in the mean of Δ=+0,255. This difference is only significant in five cases.

Table 5: Comparison of the means between the rating of the students and the auditors regarding individual items across all LSVs, leagues, and thematic blocks for the evaluation phases 1 and 2

	Evalution phase 1							Evaluation phase 2						
	Students			Auditors			Significance*	Students			Auditors			Significance*
	M	SD	n	M	SD	n		M	SD	n	M	SD	n	
Lecture														
... cardinal symptom-oriented	4,8	1,30	3308	3,8	1,77	154	**,000 (,8)**	4,9	1,31	872	4,7	1,26	36	,373 (,2)
... practice oriented	4,8	1,11	3319	4,6	1,13	153	**,027 (,2)**	5,0	1,09	872	5,1	0,83	36	,593 (,1)
... stimulates own thinking	4,3	1,29	3333	3,9	1,52	155	**,003 (,3)**	4,7	1,28	872	4,8	1,18	36	,648 (,1)
... interdisciplinary connections	3,5	1,45	3283	2,6	1,52	148	**,001 (,6)**	3,6	1,57	872	2,9	1,39	36	**,013 (,5)**
... symptom oriented knowledge	4,8	1,12	3297	4,3	1,60	151	**,002 (,4)**	4,8	1,26	872	4,6	1,17	36	,348 (,2)
... good structure	5,0	1,03	3326	4,9	1,08	154	,242 (,1)	4,8	1,25	872	4,8	1,12	36	,993 (,0)
... accordance with LSV-concept	4,6	1,29	3244	3,4	1,77	152	**,000 (,9)**	4,5	1,66	872	4,4	1,33	36	,723 (,1)
Teacher														
... prepared	5,4	0,87	3338	5,1	1,01	155	**,014 (,2)**	5,4	0,99	872	5,4	0,93	36	,991 (,0)
... friendly	5,3	0,92	3329	5,1	0,93	152	,068 (,2)	5,4	0,96	872	5,1	1,33	36	,069 (,3)
... clear wording	5,3	0,89	3335	5,1	0,94	155	**,011 (,2)**	5,2	1,06	872	5,3	0,75	36	,553 (,1)
... acoustically understandable	5,3	1,01	3337	5,0	1,14	155	**,004 (,3)**	4,9	1,39	872	5,1	1,09	36	,374 (,2)
... lively talk	4,5	1,25	3330	4,1	1,40	155	**,004 (,3)**	4,7	1,32	872	4,6	1,13	36	,546 (,1)
... interactive organisation	4,1	1,49	3328	3,7	1,71	155	**,003 (,3)**	4,6	1,36	872	4,8	1,27	36	,389 (,2)
... consideration of participants	4,8	1,11	3173	4,5	1,59	142	**,003 (,3)**	4,8	1,43	872	4,9	1,17	36	,674 (,1)
... graphic presentation	4,6	1,24	3331	4,2	1,51	155	**,003 (,3)**	4,9	1,36	872	5,0	1,16	36	,662 (,1)
... effort for learning success	4,8	1,04	3310	4,6	1,22	153	**,022 (,2)**	4,9	1,21	872	5,2	0,84	36	,081 (,3)
School grade	2,2	0,85	3247	2,6	0,97	155	**,003 (,5)**	2,1	0,81	872	2,3	1,00	36	,154 (,2)

Likert-scale: 1=strongly disagree to 6=strongly agree; * Significant p-value: p<0,05. Brackets show effect sizes (d) of differences to the mean. Bold are the items where students' and auditors' ratings differ significantly. During the first evaluation phase this is the case for almost all items. During the second evaluation phase this is the case only for one item.

of physicians and scientists lies between i_{CCmin}=-0,030 und i_{CCmax}=0,605 (see table 6). The majority of included items show a significant positive correlation. Compared to the pilot tests [32], which revealed great differences between both groups of auditors especially regarding items referring to the LSV concept, conformity between both groups is very satisfying in the first evaluation phase. In the second evaluation phase the conformity between the two auditors groups lies between i_{CCmin}=-0,022 and i_{CCmax}=0,771 and is also mostly positive significant. The conformity between the two groups of auditors can be assessed as moderately high, the intra-class correlation coefficients display a quite broad spreading.

Table 6: Intra-class correlation coefficient (with significances) for the group of the auditors – comparing students with physicians in the evaluation phases 1 and 2

	Student and physician auditors	
	$i_{CCEvaluation\ phase\ 1}$	$i_{CCEvaluation\ phase\ 2}$
Lecture		
... cardinal symptom-oriented	,504 (,000)	,709 (,000)
... practice oriented	,295 (,005)	**,245 (,151)**
... stimulates own thinking	,404 (,000)	,649 (,001)
... interdisciplinary connections	,379 (,001)	**,257 (,139)**
... symptom oriented knowledge	,433 (,000)	,383 (,049)
... good structure	,217 (,029)	,655 (,001)
... accordance with LSV-concept	,602 (,000)	,771 (,000)
Teacher		
... prepared	,394 (,000)	,751 (,000)
... friendly	**,047 (,344)**	**,021 (,463)**
... clear wording	**-,030 (,602)**	,562 (,005)
... acoustically understandable	,386 (,000)	**-,022 (,534)**
... lively talk	,484 (,000)	,487 (,015)
... interactive organisation	,605 (,000)	,592 (,003)
... consideration of participants	,479 (,000)	,517 (,010)
... graphic presentation	,539 (,000)	,737 (,000)
... effort for learning success	,374 (,000)	,497 (,014)
School grade	,548 (,000)	,620 (,002)

Bold are the items which do not show a significant correlation between student and physician auditors. These increased from two items in the first evaluation phase to four items in the second evaluation phase.

Discussion

The results of the audits and the student evaluations during the first evaluation phase reveal a picture of the LSV that overall is more positive, as expected from the results of the previous student evaluations at the end of the thematic blocks. This could indicate actual improvement. Yet it has to be taken into account that retrospective and integrated evaluations have a tendency towards worse results compared with evaluations which are performed directly after a course [31]. Hence, the observation of improvement could have been caused by a methodological effect. On the basis of school grades, the hypothetical improvement of the overall LSV rating could only be noted to a moderate degree in the second evaluation phase. The postulated criterion for improvement was only achieved in 44% of the LSVs rated by the students while an improvement in the auditor ratings was found in only 28% of the LSVs. In contrast, the auditors' ratings on the basis of individual items show mostly more positive evaluations, especially with regard to the didactic skills of the teachers in the second evaluation phase. A weakness can be seen in the small total number of lectures which is counterbalanced partly by the initial drawing of the random sample.

The chosen criterion for change – the school grade given – represents a relatively abstract measure. It can be assumed that this measure contains too little differentiation to reveal potential differences in the LSV after the intervention, since the construct "teaching quality" is a complex item [14]. The loss of information by using school grades could also be confirmed by the discrepancy within the group of assessors as far as the summative parameter of the school grade and the simultaneously rated individual items are concerned, which showed a clear improvement. During the first evaluation phase the trained assessors rated the LSV in almost all items significantly more critically compared to the participating students as was assumed in hypothesis 2. In the second evaluation phase the auditors' assessment concerning the single items turned out to be considerably better compared to the students' ratings. On one hand this could mean that an improvement of the didactic quality of the LSV had indeed taken place which was then observed and rated by the trained auditors in a more differentiated way. On the other hand the possible influence of the Rosenthal-effect must be taken into account [24], where the mere expectation of an improvement of the LSV after the intervention by the auditors could have led to a better rating. However, the assignment of trained auditors has been described as being a valid and research-oriented instrument for the rating of teaching quality [1], [17]. Others also found only moderate accordance of student and "peer-ratings" [16]. In the second evaluation phase the rating differences are less prominent which could be suggestive of a more homogenous base for the ratings according to school grade. The high inter-rater reliability hereby verified supports the validity of the data [34].

Furthermore, there is a need to analyse whether the intervention chosen for this project was potent enough to improve the LSV. Since there is no evidence in the literature that student evaluation alone improves university teaching [23], [28], an intervention beyond the mere feedback of the evaluation data was chosen for this study. Yet the feedback to the targeted group was only given in written format. Other studies show that written feedback is rarely read by the teachers and therefore may have hardly any effect [9]. More effective improvements could be reached by other interventions, e.g. didactic skill enhancing counselling [23], [33] or direct discussions with teachers about the evaluation results [2]. Feedback given as early as possible can improve the possibility of a positive effect on the teachers [26]. In our study the time between data collection and feedback was comparatively long with up to four months. On the other hand the written personal feedback was, as described in the methods section, clearly edited and illustrated in detail. It is known that written feedback of evaluations without explanations is often not correctly interpreted by the teachers and hence not understood and without effect [2]. Another influencing factor for the rather weak effect of the intervention could be down to the fact that the LSV is a multi-instructor-event with a total of approximately 150 teachers in six thematic blocks. Such a format hosts special difficulties for the realization of changes or improvements compared with courses which are taught be only a few or even a single person [26]. Furthermore it is known, that provided information or counselling of teachers in evaluation projects is less called upon if teachers are not interested or unwilling to improve their didactic skills [18]. Another important aspect for the less than dramatic effect of the intervention could be assumed to lie in the inactivity inherent to the system of faculties when it comes to the realization of curricular innovations [27]. Additionally, until the intervention during this study the concept for the LSV did not exist in a written format and was sent to the teachers during the curricular planning. With that the factor "communication within the faculty", which has a major impact during planning processes [3], was not regarded with enough attention when the new curriculum was implemented. It would be better to introduce a training procedure that acquaints all teaching personnel involved in the LSV with its concept [3]. In a subsequent survey of the quality of the LSV teachers and students should be involved to gain an acceptance as high as possible within the faculty [20]. To improve the overall effectiveness of courses the have to be integrated in a general procedure to measure and support the quality of teaching and research, since the evaluation of teaching quality alone is not sufficient for its improvement [25].

Summary and outlook

Our study demonstrated that the evaluation of the newly established LSV concept revealed didactic improvements after an intervention as well in the student ratings as in

the ratings of trained auditors. These improvements were more notable on the basis of individual items regarding the teachers or the concept rather than on the basis of school grades awarded. Students rated the LSV altogether more positive than auditors who showed a good inter-rater reliability. Apparently, a three hour training session for the auditors is not sufficient to prepare them adequately for their role as givers of analysing feedback. Furthermore, it has to be taken into account, that the generalisability of our results is somewhat reduced because of the choice of a random sample with only 18 lectures in the second evaluation phase due to the methodology chosen. The necessity of a better integration of the LSV in the global concept of the curriculum regarding content and structure with regular quality control is visible in this study. How long the effects of feedback after an evaluation last within the target group needs to be studied in further projects. Mere written information about the lecture design according to modern didactic criteria seems to be an insufficient stimulus for intervention to many teachers to improve or change their lectures. Furthermore, it needs to be checked which effects occur in the indirectly affected group of students, e.g. effects on their motivation or learning success.

Acknowledgement

We thank the Medical Faculty of Hamburg University for supporting this project (L-107/2006) from their teaching funds.

Competing interests

The authors declare that they have no competing interests.

References

1. Albanese MA, Schuldt SS, Case D, Brown D. The validity of lecturer ratings by students and trained observers. Acad Med. 1991;66(5):26-28. DOI: 10.1097/00001888-199101000-00008
2. Baggott J. Reaction of lecturers to analysis results of student ratings of their lecture skills. J Med Educ. 1987;62:491-496.
3. Bland CJ, Starnaman S, Wersal L, Moorhead-Rosenberg L, Zonia S, Henry R. Curricular change in medical schools: how to succeed. Acad Med. 2000;75(6):575-594. DOI: 10.1097/00001888-200006000-00006
4. Bortz J, Döring N. Forschungsmethoden und Evaluation. Berlin: Springer; 2006.
5. Brown G, Manogue M. AMEE Medical Education Guide No. 22: Refreshing lecturing: a guide for lecturers. Med Teach. 2001;23(3):231-244. DOI: 10.1080/01421590120043000
6. Butler JA. Use of teaching methods within the lecture format. Med Teach. 1992;14(1):11-23. DOI: 10.3109/01421599209044010
7. Cantillon P. Teaching large groups. BMJ. 2003;326:437-440.
8. Clauß G, Ebner H. Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Thun/Frankfurt a. M.: Harri Deutsch; 1977.
9. Cohen PA. Effectiveness of student-rating feedback for improving college instruction: a meta-analysis of findings. Res High Educ. 1980;13(4):321-341. DOI: 10.1007/BF00976252
10. Copeland H, Longworth D, Hewson M, Stoller J. Successful lecturing. A prospective study to validate attributes of the effective medical lecture. J Gen Intern Med. 2000;15(6):366–371. DOI: 10.1046/j.1525-1497.2000.06439.x
11. Craig M. Facilitated student discussions for evaluating teaching. SIGCSE Bulletin. 2007;39(1):190-194. DOI: 10.1145/1227504.1227376
12. Diehl JM. Normierung zweier Fragebögen zur studentischen Beurteilung von Vorlesungen und Seminaren. Psychol Erz Unterr. 2003;50:27-42.
13. Fyrenius A, Bergdahl B, Silén C. Lectures in problem-based learning - why, when and how? An example of interactive lecturing that stimulates meaningful learning. Med Teach. 2005;27(1):61-65. DOI: 10.1080/01421590400016365
14. Gordon PA. Student evaluation of college instructors: an overview. Valdosta: Valdosta State University; 1997. Zugänglich unter/available under: http://teach.valdosta.edu/WHuitt/files/tcheval.pdf
15. Grass G, Stosch C, Griebenow R. Renaissance der Vorlesung. Dtsch Ärztebl. 2005;102(23):A1642.
16. Greenwood GE, Ramagli HJ. Alternatives to student ratings of college teaching. J High Educ. 1980;51(6):673-684. DOI: 10.2307/1981172
17. Imseis HM, Galvin SL. Faculty and resident preference for two different forms of lecture evaluation. Am J Obstet Gynecol. 2004;191(5):1815-1821. DOI: 10.1016/j.ajog.2004.07.068
18. Irby D, DeMers J, Scher M, Matthews D.A model for the improvement of medical faculty lecturing. J Med Educ. 1976;51(5):403-409.
19. Leppek R, Jußen M, Berthold D, Sulzer J, Klose KJ. Windmühlenprinzip versus Uhrwerkprinzip - Tradition und Interaktion in der akademischen Vorlesung. Z Ärztl Fortbild. 1996;90:406-413.
20. Moßig I. Stichproben, Stichprobenauswahlverfahren und Berechnung des minimal erforderlichen Stichprobenumfangs. Gießen: Universität Gießen;1996.
21. Reed M. Electronic module evaluation: combining quality with quantity. Kongressbeitrag University of Leeds Inaugural Learning and Teaching Conference. Leeds: University of Leeds; 2004. Zugänglich unter/available under: http://homepages.see.leeds.ac.uk/~lecmsr/Reed%202004.doc
22. Rindermann H. Methodik und Anwendung der Lehrveranstaltungsevaluation für die Qualitätsentwicklung an Hochschulen. Sozialwis Berufspraxis. 2003;26(4):401-413.
23. Rindermann H. Quality of instruction improved by evaluation and consultation of instructors. Int J for Acad Develop. 2007;12(2):73-85. DOI: 10.1080/13601440701604849
24. Rost DH. Handwörterbuch der Pädagogischen Psychologie. Weinheim: Beltz; 2001.
25. Schmidt B. Warum oft wirksam? Und warum manchmal wirkungslos? – Subjektive Erklärungen zur Wirkung von Lehrveranstaltungsevaluation aus der Sicht von Nutzern und Anbietern. Z Eval. 2008;7(1):7-33.
26. Stillman PL, Gillers MA, Heins M, Nicholson G, Sabers D. Effect of immediate student evaluations on a multi-instructor course. J Med Educ. 1983;58:172-178.

27. Sukkar MY. Curriculum development: a strategy for change. Med Educ. 1986;20:301-306. DOI: 10.1111/j.1365-2923.1986.tb01369.x

28. Turhan K, Yaris F, Nural E. Does instructor evaluation by students using a web-based questionnaire impact instructor performance? Adv Health Sci Educ. 2005;10(1):5-13. DOI: 10.1007/s10459-004-0943-7

29. Universität Hamburg. Hamburger Lernzielkatalog. Hamburg: Universität Hamburg; 2009. Zugänglich unter/available under: http://www.uke.de/studierende/downloads/zg-studierende/Lernzielkatalog_091104_mat.pdf

30. van den Bussche H, Anders S, Ehrhardt M, Göttsche T, Hüneke B, Kohlschütter A, Kothe R, Kuhnigk O, Neuber K, Rijntjes M, Quellmann C, Harendza S. Lohnt sich eine Reform der klinischen Ausbildung? - Die Qualität des Hamburger Curriculums unter der alten und der neuen Approbationsordnung im Vergleich. Z Ärztl Fortbild Qualitätssich. 2005;99:419-423.

31. van den Bussche H, Weidtmann K, Kohler N, Frost M, Kaduskiewicz H. Evaluation der ärztlichen Ausbildung: Methodische Probleme der Durchführung und der Interpretation von Ergebnissen. GMS Z Med Ausbild. 2006;23(2):Doc37. Zugänglich unter/available under: http://www.egms.de/de/journals/zma/2006-23/zma000256.shtml

32. Weidtmann K. Analyse des Status quo der Leitsymptom-Vorlesung und Planung einer evaluationsbasierten Intervention an der Medizinischen Fakultät Hamburg. Unveröffentlichte Projektarbeit im Studiengang Master of Medical Education. Heidelberg: Medizinische Fakultät Heidelberg; 2007.

33. Wilson RC. Improving faculty teaching: Effective use of student evaluations and consultants. J High Educ. 1986;57(2):196-211. DOI: 10.2307/1981481

34. Wirtz M. Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen. Rehabilitation. 2004;43:384-389. DOI: 10.1055/s-2003-814935

Corresponding author:
Dr. med. Olaf Kuhnigk, MME (Bern)
Universitätsklinikum Hamburg-Eppendorf, Klinik für Psychiatrie und Psychotherapie, Martinistraße 52, 20246 Hamburg, Deutschland, Tel.: +49 (0)40/7410-57675, Fax: +49 (0)40/7410-54702
o.kuhnigk@uke.de

Please cite as
Kuhnigk O, Weidtmann K, Anders S, Hüneke B, Santer R, Harendza S. Leitsymptomvorlesungen im klinischen Studienabschnitt - Effekte evaluationsbasierter Interventionen auf eine Großgruppen-Lehrveranstaltung. GMS Z Med Ausbild. 2011;28(1):Doc15.
DOI: 10.3205/zma000727, URN: urn:nbn:de:0183-zma0007272

This article is freely available from
http://www.egms.de/en/journals/zma/2011-28/zma000727.shtml

Received: *2010-01-15*
Revised: *2010-08-02*
Accepted: *2010-09-23*
Published: *2011-01-04*