

## » Entwicklung eines Fragebogens zur Evaluation von makroskopischen Präparierübungen

H. P. Huber<sup>1</sup>, F. Anderhuber<sup>2</sup>

<sup>1</sup> Abteilung für Klinische Psychologie, Institut für Psychologie, Karl-Franzens-Universität Graz

<sup>2</sup> Anatomisches Institut, Karl-Franzens-Universität Graz

**Zusammenfassung. Ziele:** Die vorliegende Studie hatte sich zum Ziel gesetzt, nach den Prinzipien der klassischen Testtheorie einen psychometrisch fundierten Fragebogen zu entwickeln, der zur Evaluation von makroskopischen Präparierübungen durch Studierende geeignet sein sollte. **Methodik:** Die Testentwicklung von der Rohfassung bis zur ersten Erprobung der vorliegenden Fragebogen-Version erfolgte an drei Stichproben ( $N_1 = 217$ ,  $N_2 = 50$ ,  $N_3 = 326$ ) von Studierenden der Medizin im vorklinischen Studienabschnitt. Die Itemselektion orientierte sich an den Kriterien der Schwierigkeit und Trennschärfe sowie am Beitrag der Items zur internen Konsistenz. **Ergebnisse:** Aufgrund der Itemanalyse ergab sich eine (vorläufige) Endfassung, die aus 30 Items (Aussagen, Feststellungen) besteht. Diese gliedern sich in vier Untergruppen (Subtests), die die Bereiche „Kursinhalt und Ablauf“, „übungsbegleitende Faktoren“, „Lehrpersonal“ und „Relevanz und Zufriedenheit“ abdecken. Die Reliabilitäten (interne Konsistenz) der drei ersten Subtests (Kernskalen) variieren zwischen 0,60 und 0,79. Die Resultate einer ersten Erprobung bestätigten die Brauchbarkeit des Fragebogens. **Schlussfolgerung:** Der GFEP („Grazer Fragebogen zur Evaluation von makroskopischen Präparierübungen“) ermöglicht eine standardisierte Evaluation der Kursgestaltung sowie des Unterrichtsverhaltens und der didaktischen Kompetenz des Lehrpersonals durch Studierende. Zweck der Evaluation ist die Verbesserung der Lehre und die Optimierung der Kursgestaltung.

**Development of a Questionnaire for Evaluating Dissection Courses. Purpose:** Following the principles of classical test theory, an attempt was made to develop a standardised questionnaire for students' evaluations of dissection courses. **Method:** Questionnaire data of three samples ( $N_1 = 217$ ,  $N_2 = 50$ ,  $N_3 = 326$ ) of medical students were subjected to classical item analyses, including computation of item difficulty, item-test correlation and coefficient  $\alpha$ . The scheme of item scoring utilised a grade mark scale ranging from 1 to 5. **Results:** Item analyses resulted in a questionnaire comprising 30 items. The test items covered four domains (subtests): „Course Content and Course Implementation“, „Course Quality“, „Teachers' Competence“ and „Utility and General Satisfaction“. Coefficients of reliability esti-

mated in terms of Cronbach's  $\alpha$  ranged from 0.60 to 0.79 for the first three subtests. A first test of application provided useful results. **Conclusions:** The questionnaire (GFEP – Grazer Fragebogen zur Evaluation von makroskopischen Präparierübungen) provides sufficiently reliable class average student ratings for a multidimensional evaluation of dissection courses. The evaluation results are expected to contribute to the improvement of teaching and course structure.

**Key words:** Evaluation – Dissection courses – Classical test theory – Item analysis

### Fragestellung

Mit dem Begriff „Evaluation“ werden sehr unterschiedliche Vorstellungen assoziiert. Dazu gehören beispielsweise Konzepte wie Produktbewertung, Qualitätskontrolle, Prozessbeurteilung, Programmbewertung oder Erfolgskontrolle. Evaluation ist nicht Selbstzweck, sondern dient der Handlungsoptimierung. Von den Ergebnissen einer Evaluation erwartet man sich vor allem Hilfe bei der Planung, Überprüfung und Verbesserung praktischer Maßnahmen. Je nach den Rahmenbedingungen, den Grundsatzzielen, der Ausrichtung der Fragestellung, dem Zeitpunkt, den zugrunde liegenden Kosten-Nutzen-Überlegungen und der Bearbeitungsform eines Evaluationsprojektes lassen sich nach Wottawa u. Thierau [1] u. a. folgende Evaluationsformen unterscheiden: a) offene und geschlossene Evaluationen, b) praxis-, entwicklungs- und theorieorientierte Evaluationen, c) Mikro- und Makroevaluationen, d) prognostisch-prospektive, dynamische und ergebnisorientierte Evaluationen, e) strategische und managementorientierte Evaluationen, f) Selbst- und Fremdevaluationen sowie g) vergleichende, summative (zusammenfassende) und formative (auf Gestaltungsverbesserung abzielende) Evaluationen. Unter Metaevaluation versteht man die Aufarbeitung und Zusammenfassung der nach Abschluss eines Evaluationsprojektes gemachten Erfahrungen.

Im internationalen Vergleich befindet sich die Evaluation der Lehre an den deutschsprachigen Universitäten noch im Anfangsstadium. Während sich in Großbritannien 95%, in den USA 89%, in den Niederlanden 88% und in Australien 77% der Professoren einer regelmäßigen Evaluation unterziehen, liegt Deutschland mit 18% deutlich unter diesen Werten [2]. In

Österreich beginnt die Evaluation von Lehrveranstaltungen überhaupt erst Fuß zu fassen [3].

Im Folgenden soll ein nach den Prinzipien der klassischen Testtheorie [4] entwickelter Fragebogen vorgestellt werden, der zur Evaluation von makroskopischen Präparierübungen konzipiert wurde. Es handelt sich dabei nicht um eine Fremdbeurteilung durch Außenstehende, sondern um eine *interne Bewertung*, die von den *Studierenden* vorgenommen wird. Der „Grazer Fragebogen zur Evaluation von makroskopischen Präparierübungen“ (GFEP) versteht sich als ein ergebnisorientiertes Rückmeldeverfahren mit vorwiegend summativem Charakter. „Summativ“ heißt, dass die Studierenden am Ende des Semesters gebeten werden, anhand der Fragebogenitems eine zusammenfassende Globalbeurteilung der absolvierten Lehrveranstaltung abzugeben. Die mit dem GFEP vorgenommene Evaluation verfolgt drei Ziele: a) den Kursinhalt sowie die organisatorischen Rahmenbedingungen der Präparierübungen zu optimieren, b) den Lehrveranstaltungsleitern Hinweise zur Verbesserung ihrer didaktischen Bemühungen zu geben und c) Informationen über die Akzeptanz der Lehrveranstaltung bei den Studierenden einzuholen.

Im nächsten Abschnitt sollen zunächst das Material, die Methoden und die Phasen der Testentwicklung kurz beschrieben werden. Die Konstruktionsphasen I und II wurden von Frau Hörtner (1998) im Rahmen ihrer Diplomarbeit durchgeführt. Im Ergebnisteil wird zum einen die vorläufige Endform des Fragebogens vorgestellt, zum anderen wird über das Ergebnis einer ersten Anwendung berichtet.

## Methodik

### Material

Der Itempool wurde an der Karl-Franzens-Universität Graz vom Anatomischen Institut in Zusammenarbeit mit dem Institut für Psychologie erstellt. Als Informationsquellen für die Itemgenerierung dienten a) Anregungen von Studierenden aus Lehrveranstaltungskritiken, b) das handlungsrelevante Wissen der Lehrveranstaltungsleiter des Anatomischen Instituts sowie c) Annahmen und Konzepte aus der hochschuldidaktischen Fachliteratur.

Die Items der Rohfassung gliederten sich in vier Untergruppen („Subtests“), die verschiedene Beurteilungsaspekte abdecken sollten. Der erste Untertest (13 Items) bezieht sich auf den „Kursinhalt und Ablauf“ (KA); er thematisiert u. a. die Gestaltung und den Inhalt der Übungen, den Wert der Gruppenarbeit sowie die Relevanz für das Rigorosum und die spätere Berufspraxis. Der Subtest „Übungsbegleitende Faktoren“ (ÜF, sechs Items) beinhaltet Aussagen über den Bezug der Vorlesungen zu den Übungen, die Nützlichkeit der angegebenen Lernunterlagen und die Qualität der Videodemonstrationen. Der Untertest „Lehrpersonal“ (LP, acht Items) wurde als „Dozentenskala“ konzipiert, die eine Beurteilung der fachlichen Kompetenz und des didaktischen Engagements der Lehrveranstaltungsleiter und Tutoren ermöglichen soll. Mit dem vierten Untertest „Relevanz und Zufriedenheit“ (RZ, fünf Items) soll eine globale Einschätzung der Kursatmosphäre („soziales Klima“) sowie der Bedeutung der Übungen für die medizinische Ausbildung erfolgen. Das Ausfüllen des Fragebogens dauert etwa zehn Minuten.

### Zur Methode der Testkonstruktion

Die Itemanalyse umfasste die Berechnung der Schwierigkeitsindizes und Trennschärfekoeffizienten [4]. Der Schwierigkeits- bzw. Popularitätsindex  $p_j$  gibt die relative Häufigkeit jener Personen an, die ein Item  $j$  in der Schlüsselrichtung beantwortet haben. Mit  $p_j = 0,5$  erreicht ein Item sein Differenzierungsoptimum. Der Trennschärfekoeffizient  $r_{jt}$  eines Items  $j$  wurde als Produkt-Moment-Korrelation zwischen den Itempunktwerten und den Rohwerten definiert. Um Scheinkorrelationen (part-whole correlations) zu vermeiden, blieb bei der Rohwertbildung jeweils der Punktwert jenes Items unberücksichtigt, für welches der Trennschärfekoeffizient berechnet wurde. Die Reliabilitätsschätzungen der Subtests erfolgten konsistenzanalytisch. Der hier verwendete  $\alpha$ -Koeffizient von Cronbach stellt eine Maximum-Likelihood-Schätzung der Reliabilität dar.

### Phase I der Testentwicklung

Die Rohfassung wurde 74 männlichen und 143 weiblichen Studierenden der Medizin im Anschluss an die Beendigung des Präparierkurses vorgegeben. Das Alter der Befragten streute um einen Mittelwert von  $M = 20,7$  zwischen 18 und 54 Jahren. Die mittlere Studiendauer betrug vier Semester. Eine Zusammenfassung der Ergebnisse der Itemanalyse und der Reliabilitätsbestimmungen findet man in Tab. 1.

**Tab. 1** Vergleich der Rohfassung (Phase I) mit der revidierten Version (Phase II) hinsichtlich der mittleren Schwierigkeitsindizes ( $p$ ), der mittleren Trennschärfekoeffizienten ( $r_{jt}$ ) und der inneren Konsistenz ( $\alpha$ -Koeffizienten nach Cronbach)

Subtests	Rohfassung			Revision		
	mittl. $p$	mittl. $r_{jt}$	$\alpha$	mittl. $p$	mittl. $r_{jt}$	$\alpha$
KA	0,41	0,10	0,35	0,34	0,33	0,63
ÜF	0,37	0,23	0,45	0,35	0,45	0,73
LP	0,43	0,44	0,73	0,47	0,46	0,78
RZ	0,25	0,24	0,41	0,32	0,28	0,52

Anmerkungen. Subtests: KA = Kursinhalt und Ablauf, ÜF = übungsbegleitende Faktoren, LP = Lehrpersonal, RZ = Relevanz und Zufriedenheit.

### Phase II der Testentwicklung

Die itemanalytischen Befunde machten eine Reihe von Modifikationen notwendig. Ein Änderungsbedarf war vor allem in Hinblick auf die Trennschärfekoeffizienten bei den Subtests KA und RZ gegeben. Die wichtigste Modifikation bestand darin, dass für alle Items in Anlehnung an die vertraute Notenskala ein fünfstufiges Antwortformat (1 = sehr gut, 5 = sehr schlecht) eingeführt wurde. Die Änderung des Antwortformats hatte in der Regel auch eine Änderung der Itemformulierung zur Folge. Dadurch erhöhte sich im Subtest LP die Anzahl der Items von 8 auf 9.

In einem Probelauf wurde dann die revidierte Fassung einer Stichprobe von  $N = 50$  Studierenden (25 Männer und 25 Frauen) vorgegeben. Wie Tab. 1 zeigt, konnte durch die in Phase II vorgenommene Revision in allen Subtests eine Verbesserung der Trennschärfe und Reliabilität erzielt werden.

**Ergebnisse**

Als Resultat der Testkonstruktion kann die im Anhang (Abb. 1) wiedergegebene vorläufige Endform des „Grazer Fragebogens zur Evaluation von makroskopischen Präparierübungen“ (GFEP) vorgelegt werden. Sie umfasst insgesamt 30 Items, da drei Items wegen unzureichender Trennschärfe ausgeschieden wurden. Die Konstruktionsdaten sind in Tab. 2 zusammengefasst. Die Interkorrelationen zwischen den Subtests bewegen sich zwischen 0,33 und 0,47 ( $p < 0,01$  für alle  $r$ -Werte).

**Tab. 2** Mittlere Itemwerte ( $M$ ), Standardabweichungen ( $s$ ), mittlere Schwierigkeit ( $p$ ) und Trennschärfe ( $r_{jt}$ ) sowie innere Konsistenz ( $\alpha$ -Koeffizient nach Cronbach)

Subtests	$M$	$s$	mittl. $p$	mittl. $r_{jt}$	$\alpha$
Kursinhalt und Ablauf (KA)	1,99	1,37	0,25	0,35	0,66
übungsbegleitende Faktoren (ÜF)	2,27	1,45	0,32	0,32	0,60
Lehrpersonal (LP)	2,38	1,90	0,34	0,49	0,79
Relevanz und Zufriedenheit (RZ)	2,23	1,05	0,31	0,24	0,48

Außer den bereits besprochenen Subtests enthält der Fragebogen noch Teile, die keiner Itemanalyse unterzogen wurden. Dabei handelt es sich um a) Angaben zur eigenen Person (wie z. B. Alter, Geschlecht, Semesterzahl, sezierte Region, angestrebte Fachrichtung), b) ergänzende Fragen zu bestimmten Items, die dann zu beantworten sind, wenn der durch diese Items angesprochene Sachverhalt (z. B. „Ausgewogenheit zwischen Theorie und Praxis“) mit „schlecht“ oder „sehr schlecht“ bewertet wurde, und c) Antwortkategorien, die bewusst auf ein gebundenes Antwortformat verzichten, um Beurteilungen in eigenen Worten zu ermöglichen (z. B. „Was mir besonders gefällt“, „Was mich stört“, „Verbesserungsvorschläge“).

Der erste Einsatz des Fragebogens erfolgte an einer Stichprobe von  $N = 326$  Studierenden (183 Frauen, 143 Männer), die gerade den makroskopischen Präparierkurs für Fortgeschrittene absolviert hatten. Die Rücklaufquote betrug 100%. Die Ergebnisse sind in Tab. 2 zusammengefasst. Wie die mittleren Itemwerte (Notenskala) für die einzelnen Subtests zeigen, wurden sowohl die Gestaltung und der Inhalt der Übungen (Subtest KA) als auch die Nützlichkeit der angegebenen Lernunterlagen (Subtest ÜF) als auch die fachliche Kompetenz des Lehrpersonals und die Betreuung (Subtest LP) sowie die Relevanz der Übung für die Praxis und die Kursatmosphäre (Subtest RZ) von den Studierenden als „gut“ eingestuft.

Spezifischere Informationen erhält man, wenn man die Itemebene betrachtet. Hier zeigte sich u. a., dass das Item 25 („Ausmaß der Anregungen zur Diskussion“) mit  $M = 3,37$  ( $s = 1,17$ ) am schlechtesten beurteilt wurde. Die beste Beurteilung erhielt mit  $M = 1,30$  ( $s = 0,56$ ) das Item 10 („Bezug der Vorlesungen zu den Übungen“). Ferner ergab ein Geschlechtervergleich mit dem U-Test von Mann und Whitney [5], dass die Angemessenheit der täglichen Übungszeit ( $u = -2,65$ ,  $p = 0,008$ ), die Qualität der Videodemonstrationen ( $u = -2,04$ ,  $p = 0,04$ ) und die Möglichkeit, Fragen zu stellen ( $u = -2,44$ ,  $p = 0,02$ ), von

**ANGABEN ZUR EIGENEN PERSON**

- Geschlecht  männlich     weiblich
- Alter in Jahren .....
- Semesterzahl .....
- präparierte Region  Kopf-Hals     Brust     Bauch
- Berufsziel - angestrebte Fachrichtung
- Facharzt für Allgemeinmedizin
- Facharzt für chirurgisches/morphologisches Fach inkl. Orthopädie, Gynäkologie, HNO, Urologie, Anästhesie, Gerichtsmedizin, Pathologie, Radiologie, Neurologie
- Facharzt für ein anderes Fach  
.....

**A. KURSINHALT UND ABLAUF**

<b>1. Ausgewogenheit zwischen Theorie und Praxis</b>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> sehr gut <span style="float: right;">sehr schlecht</span>
Bei Beurteilung "schlecht" oder "sehr schlecht" a) zu viel Praxis <span style="float: right;">b) zu viel Theorie</span> <input type="checkbox"/> <span style="float: right;"><input type="checkbox"/></span>	
<b>2. Angemessenheit des Inhalts der Übungen</b>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> sehr gut <span style="float: right;">sehr schlecht</span>
Bei Beurteilung "schlecht" oder "sehr schlecht" a) zu wenig <span style="float: right;">b) zu viel</span> <input type="checkbox"/> <span style="float: right;"><input type="checkbox"/></span>	
<b>3. Angemessenheit der Gesamtübungszeit</b>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> sehr gut <span style="float: right;">sehr schlecht</span>
Bei Beurteilung "schlecht" oder "sehr schlecht" a) zu kurz <span style="float: right;">b) zu lange</span> <input type="checkbox"/> <span style="float: right;"><input type="checkbox"/></span>	
<b>4. Angemessenheit der täglichen Übungszeit</b>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> sehr gut <span style="float: right;">sehr schlecht</span>
Bei Beurteilung "schlecht" oder "sehr schlecht" a) zu kurz <span style="float: right;">b) zu lange</span> <input type="checkbox"/> <span style="float: right;"><input type="checkbox"/></span>	
<b>5. Regioneneinteilung (zugeteilte Körperregion)</b>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> sehr gut <span style="float: right;">sehr schlecht</span>
<b>6. Wert der Gruppenarbeit</b>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> sehr gut <span style="float: right;">sehr schlecht</span>

**Abb. 1** Anhang: „Grazer Fragebogen zur Evaluation von makroskopischen Präparierübungen“ (GFEP)

7. Angemessenheit der Anforderungen (hinsichtlich des vorausgesetzten Wissensniveaus)	<input type="checkbox"/>				
	sehr gut				sehr schlecht

Bei Beurteilung "schlecht" oder "sehr schlecht"  
 a) zu hoch  b) zu niedrig

8. Angemessenheit der Testurenanzahl	<input type="checkbox"/>				
	sehr gut				sehr schlecht

Bei Beurteilung "schlecht" oder "sehr schlecht"  
 a) zu wenig  b) zu viel

9. Organisation insgesamt	<input type="checkbox"/>				
	sehr gut				sehr schlecht

**B. ÜBUNGSBEGLEITENDE FAKTOREN**

10. Bezug der Vorlesungen zu den Übungen	<input type="checkbox"/>				
	sehr gut				sehr schlecht

11. Kombination der Vorlesung mit Gruppenunterricht in den Übungen	<input type="checkbox"/>				
	sehr gut				sehr schlecht

12. Nützlichkeit der angegebenen Lernunterlagen	<input type="checkbox"/>				
	sehr gut				sehr schlecht

13. Wahlmöglichkeit der Prüfer	<input type="checkbox"/>				
	sehr gut				sehr schlecht

14. Nützlichkeit der Videodemonstration	<input type="checkbox"/>				
	sehr gut				sehr schlecht

15. Dauer der Videodemonstration	<input type="checkbox"/>				
	sehr gut				sehr schlecht

Bei Beurteilung "schlecht" oder "sehr schlecht"  
 a) zu lange  b) zu kurz

16. Qualität der Videodemonstration	<input type="checkbox"/>				
	sehr gut				sehr schlecht

**C. LEHRPERSONAL**

17. Die Anzahl des Lehrpersonals ist ausreichend	<input type="checkbox"/>				
	sehr gut				sehr schlecht

18. Anleitung (z. B. Anregungen zu den Überlegungen eines Präparationsschrittes) durch den zuständigen Demonstrator	<input type="checkbox"/>				
	sehr gut				sehr schlecht

19. Betreuung durch das weitere Lehrpersonal (supervidierende Professoren und Assistenten)	<input type="checkbox"/>				
	sehr gut				sehr schlecht

20. Ausmaß der praktischen Hilfestellung bei der Präparation	<input type="checkbox"/>				
	sehr gut				sehr schlecht

Bei Beurteilung "schlecht" oder "sehr schlecht"  
 a) zu viel  b) zu wenig

21. Ausmaß der Gelegenheit, Fragen zu stellen	<input type="checkbox"/>				
	sehr gut				sehr schlecht

22. Fachliche Kompetenz des Lehrpersonals	<input type="checkbox"/>				
	sehr gut				sehr schlecht

23. Verständlichkeit der Stoffvermittlung durch das Lehrpersonal	<input type="checkbox"/>				
	sehr gut				sehr schlecht

24. Fachübergreifende Stoffvermittlung (z. B. Hinweise auf die praktische Anwendbarkeit innerhalb des Medizinstudiums)	<input type="checkbox"/>				
	sehr gut				sehr schlecht

25. Ausmaß der Anregung zur Diskussion	<input type="checkbox"/>				
	sehr gut				sehr schlecht

**D. RELEVANZ UND ZUFRIEDENHEIT**

26. Relevanz der Übungen für das Rigorosum	<input type="checkbox"/>				
	sehr gut				sehr schlecht

27. Relevanz der Übungen für die Praxis/den Beruf	<input type="checkbox"/>				
	sehr gut				sehr schlecht

28. Angemessenheit der Zugangskriterien	<input type="checkbox"/>				
	sehr gut				sehr schlecht

Bei Beurteilung "schlecht" oder "sehr schlecht"  
 a) zu viel  b) zu wenig

29. Allgemeine Stimmung (Kursatmosphäre, soziales Klima)	<input type="checkbox"/>				
	sehr gut				sehr schlecht

30. Gesamtbeurteilung	<input type="checkbox"/>				
	sehr gut				sehr schlecht

Fortsetzung von Abb. 1

Fortsetzung von Abb. 1

**Was mir besonders**

gefällt:.....  
 .....  
 .....

**Was mich besonders**

stört:.....  
 .....  
 .....

**Was ich anders machen würde, wenn ich die Sezlerübungen nochmals machen**

könnte:.....  
 .....  
 .....  
 .....

**Verbesserungsvorschläge:**.....  
 .....  
 .....  
 .....

Fortsetzung von Abb. 1

den Studenten besser beurteilt wurde als von den Studentinnen.

**Diskussion**

Der vorliegende Fragebogen wurde weder zur Selbst-Evaluation der Lehrenden, noch zur Fremd-Evaluation durch externe Experten entwickelt. Er ist vielmehr ein Instrument, das eine *standardisierte Evaluation* der Präparierübungen durch die *Studierenden* ermöglichen soll. Das Kernstück des „Grazer Fragebogens zur Evaluation von makroskopischen Präparierübungen“ sind die Subtests (Skalen) „Kursinhalt und Ablauf“ (KA), „Übungsbegleitende Faktoren“ (ÜF) und „Lehrpersonal“ (LP). Sie weisen eine hinreichende Reliabilität auf, wenn die Mittelwerte der von den Studierenden abgegebenen Beurteilungen betrachtet werden. Die Ergebnisse einer ersten Erprobung an einer größeren Stichprobe haben gezeigt, dass der Fragebogen sowohl zur Modifikation des formalen Übungsablaufs und als auch zur Verbesserung der didaktischen Kursgestaltung wertvolle Informationen liefern kann. Eine noch zu korrigierende Schwachstelle ist zweifelsohne der Subtest „Relevanz und Zufriedenheit“ (RZ).

Bisweilen wird bezweifelt, ob Studierende überhaupt in der Lage sind, das Geschehen einer Lehrveranstaltung adäquat zu beschreiben. Diese Skepsis ist jedoch weitgehend unbegründet [6]. Nach Marsh ([7], S. 707) sind die Durchschnittswerte der von den Studierenden abgegebenen Beurteilungen („class-average students ratings“) multidimensionale, stabile und reliable Indikatoren der Lehrkompetenz und des didaktischen Engagements des Veranstaltungleiters, die von möglichen Biasvariablen (etwa vom veranstaltungsspezifischen Studienerfolg) weitgehend unbeeinflusst bleiben.

Grundsätzlich ist festzuhalten, dass Evaluationen nur dann den damit verbundenen Aufwand rechtfertigen, wenn sie auch Konsequenzen nach sich ziehen. Gegebenenfalls sind für das evaluierte Lehrpersonal entsprechende Beratungs- und hochschuldidaktische Weiterbildungsangebote vorzusehen [8].

**Literatur**

- <sup>1</sup> Wottawa H, Thierau H. Lehrbuch Evaluation (2. Aufl.). Bern: Huber; 1998
- <sup>2</sup> Rindermann H. Die studentische Beurteilung von Lehrveranstaltungen: Forschungsstand und Implikationen für den Einsatz von Lehrevaluationen, Tests und Trends 11. Jahrbuch der Pädagogischen Diagnostik. (im Druck)
- <sup>3</sup> Sonneck G. Bildungsziele und Lehrveranstaltungen im Medizinstudium. Mitteilungen der Studienkommission Wien. Wien: Facultas; 1994
- <sup>4</sup> Lienert GA, Raatz U. Testaufbau und Testanalyse. 5. Auflage. Weinheim: Psychologie VerlagsUnion; 1994
- <sup>5</sup> Bortz J, Lienert GA, Boehnke K. Verteilungsfreie Methoden in der Biostatistik. Berlin: Springer; 1990
- <sup>6</sup> Rindermann H. Zur Qualität studentischer Lehrveranstaltungsevaluationen: Eine Antwort auf Kritik an der Lehrevaluation. Zschr Pädag Psychol 1996; 10: 129 – 145
- <sup>7</sup> Marsh HW. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential bias, and utility. J Educ Psychol 1984; 76: 707 – 754
- <sup>8</sup> Wilson RC. Improving faculty teaching. Effective use of student evaluations and consultants. J Higher Educ 1986; 57: 196 – 211

Univ.-Prof. Dr. Helmuth P. Huber

Institut für Psychologie  
 Karl-Franzens-Universität Graz  
 Universitätsplatz 2/III  
 8010 Graz  
 Österreich

E-mail: helmuth.huber@kfunigraz.ac.at