

Der Einfluss von Anonymität in der Lehrevaluation durch Studierende

Zusammenfassung

Zielsetzung und Fragestellung: Unterscheiden sich die Ergebnisse anonymisierter und personalisierter Evaluationen?

Methodik: Während zweier Jahre wurden bei jeder Modulevaluation alle Studierenden randomisiert den Halbgruppen „anonym“ und „personalisiert“ zugeteilt. Die Qualität des Moduls in seinen relevanten Aspekten wurde mit einem standardisierten Fragebogen erhoben. Zusätzlich konnten optionale Textantworten formuliert werden. Drei unabhängige Personen bewerteten die Aussagequalität der Antworten gemäss einem Leitfaden. Dieser beinhaltete die fünf Dimensionen positiv-negativ, differenziert-absolut, Nennung einer Person-allgemein, Befehl enthaltend-neutral und optisch akzentuiert-blank.

Der Datensatz bestand aus 615 Fragebogen, davon waren 306 in anonymer Form. Mit einer multivariaten Varianzanalyse wurde überprüft, ob sich zwischen anonym und personalisiert erhobenen Daten ein Unterschied sowohl bei den durch Skalen bewerteten Aspekten als auch bei der Qualität der optionalen Textantworten zeigte. Zusätzlich wurde untersucht, ob sich die Häufigkeit der optionalen Textantworten unterschied.

Ergebnisse: In der Aussagequalität liess sich kein signifikanter Unterschied zwischen anonym und personalisiert erhobenen Daten nachweisen. Hingegen haben Studierende aus der personalisierten Gruppe tendenziell häufiger optionale Textantworten geliefert.

Schlussfolgerung: Personalisierte Evaluationen generieren dann keine verfälschten Resultate im Sinne der sozialen Erwünschtheit, wenn der Evaluationszirkel geschlossen und transparent ist: Die Verantwortlichen melden den Studierenden die Evaluationsergebnisse konsequent zurück und Änderungs-, Optimierungswünsche und Umsetzungsmöglichkeiten werden mit ihnen diskutiert. Die Studierenden erfahren so, dass ihre Rückmeldungen ernst genommen werden. So können sie sich echte Kritik erlauben.

Schlüsselwörter: Ausbildung, Curriculumentwicklung, Programmevaluation, Anonymität der Befragten

Einleitung

Ausgangslage

Der Bachelorstudiengang der Berner Fachhochschule (BFH) in Pflege wurde im Jahr 2006 gemäss den Vorgaben der Bologna-Reform neu entwickelt. Der Lehrgang ist als kompetenzbasiertes Generalistenstudium angelegt, das wissenschaftliche Fundierung mit konsequenter Praxisorientierung verbindet. Die Struktur ist modular, das pädagogisch-didaktische Konzept ist das Problembasierte Lernen (PBL: Problem-based Learning). Das Kollegium besteht aus zwanzig Dozierenden, die in der Regel eine Grundausbildung in Pflege und einen akademischen Abschluss in Pflegewissenschaft, in Erziehungswissenschaft oder Psychologie mitbringen. Pro Studienjahr werden im Schnitt 100 Pflegende ausgebildet.

Der Evaluationsprozess des Studiengangs wurde auf der Basis der Evaluationsliteratur, die sich für PBL-Curricula bewährt hat, parallel zum Curriculum entwickelt [1], [2]. Das Ziel der Lehrevaluation (Modulevaluationen, Dozierendenevaluationen, etc.) war die kontinuierliche Optimierung des Studiengangs. Daraus resultierte ein entsprechendes Evaluationskonzept, das sich folgendermassen charakterisieren lässt:

1. **Kontinuierliche Evaluation aller Module durch Studierende und Dozierende** für strukturell-organisatorische Anpassungen innerhalb der einzelnen Module.
2. **Kontinuierliche Anpassung des Curriculums durch die Surveygruppe.** Dies ist die Evaluationsgruppe, bestehend aus der Studiengangsleiterin, dem Leiter Ressort Ausbildungsprogramm und dem wissenschaft-

Theresa Scherer¹

Jan Straub¹

Daniel Schnyder¹

Noemi Schaffner¹

1 Berner Fachhochschule,
Studiengang Pflege, Bern,
Schweiz

lichem Mitarbeiter. Für strukturell-inhaltliche Anpassungen innerhalb des gesamten Curriculums

3. **Kontinuierliche Qualitätsentwicklung durch den Qualitätszirkel.** Dieser besteht aus dem gesamten Kollegium. Der Qualitätszirkel sorgt für Anpassungen systemisch-haltungsmässiger Art, z. B. betreffend die pädagogische Grundhaltung, für die Realisation des zukünftigen Berufsbildes.

Die Modulevaluation durch die Studierenden geschah mit einem standardisierten Fragebogen am Ende jedes Moduls. Die Fragen zielten darauf ab, die Qualität der Module zu überprüfen. Die Studierenden füllten den Fragebogen jeweils im Rahmen einer Lehrveranstaltung am Ende des Moduls aus. Dadurch wurde mit über 95 Prozent eine sehr hohe Rücklaufquote erreicht. Alle Evaluationen fanden anonymisiert statt.

Die erhobenen Daten wurden statistisch ausgewertet und in einer Zusammenfassung dargestellt. Diese Zusammenstellung wurde den Studierenden im darauffolgenden Modul durch die Studiengangsleitung präsentiert; die Ergebnisse wurden diskutiert; zu Änderungswünschen oder Anregungen der Studierenden erfolgte eine Rückmeldung darüber, ob und falls ja, wie die Vorschläge umgesetzt würden.

Dieses Vorgehen bewährte sich in den vergangenen Jahren. Hingegen löste die Frage, ob die Studierendenbefragung weiterhin anonym durchgeführt werden sollte immer wieder kontroverse Diskussionen im Kollegium aus. Anlass dazu waren einzelne numerische Bewertungen, die auffallend schlecht waren, z. B. durchwegs Note 1¹ oder Kommentare, die vernichtend und verletzend formuliert waren, z. B. „Frau Müller² ist vollkommen unfähig! Schicken Sie sie in eine Weiterbildung!“ Zudem stimmten sehr negativ formulierte Textantworten der Studierenden mit den dazugehörigen numerischen Bewertungen oft nicht überein. In solchen Fällen konnte das Kollegium nicht nachvollziehen, wie die Ergebnisse interpretiert werden sollten. Da die Studierenden die Module ohne Namensangabe evaluierten, konnten die Zuständigen nicht nachfragen, was zu grossen Irritationen führte. Ein Teil des Kollegiums vermutete, dass ein anonymes Evaluieren diese sehr negativen Studierendenrückmeldungen fördert. Die zu diesem Zeitpunkt vorhandene Datenmenge war allerdings sehr gering, weshalb dazu keine fundierten Rückschlüsse gezogen werden konnten. Ein Argument der Befürworter des anonymisierten Vorgehens war die Vermutung, Studierende könnten bei negativer Kritik belastende Konsequenzen von Seiten der Dozierenden befürchten. Auch in der Forschungsliteratur wird darauf hingewiesen, dass insbesondere bei sensiblen oder bedrohlichen Antworten, die negative Auswirkungen haben könnten, die Gefahr besteht, dass der Fragebogen nicht ehrlich, sondern im Sinne von sozialer Erwünschtheit ausgefüllt wird [3]. Zudem wiesen verschiedene Studien nach, dass Befragte dazu tendieren, gefälliger Antworten zu geben, wenn sie wissen oder vermuten, dass ihre Antworten zusammen mit ihrem Namen von Dozierenden eingesehen werden können [4], [5]. Gemäss einer ande-

ren Untersuchung von Fries und McNinch [6] notieren Studierende ihren Namen nicht, wenn sie etwas Negatives zu sagen haben, selbst dann nicht, wenn sie zur Namensnennung aufgefordert worden sind. Aus diesen Gründen empfiehlt die Fachliteratur, Fragebogen und Beurteilungsskalen anonym ausfüllen zu lassen, um die Genauigkeit und die Verlässlichkeit der Daten zu gewährleisten [3]. Aber ebenso viele Untersuchungen konnten keine Unterschiede zwischen anonym und personalisiert erhobenen Daten nachweisen [7], [8]. Zudem nimmt das Kollegium gemäss dem pädagogischen Grundkonzept die Haltung ein, dass die Studierenden als erwachsene Partnerinnen und Partner betrachtet werden sollen und dass ein sachlicher, offener Austausch bezüglich der Leistungsqualität möglich sein muss. Die Studierenden sollten im Laufe ihres Studiums dazu hingeführt werden, kritische, aber respektvolle und verantwortungsvolle Rückmeldungen geben zu können.

Fragestellung

Bisherige Forschungsergebnisse zu anonymen bzw. personalisierter Evaluation sind widersprüchlich und die Fachliteratur ist teilweise älter als zehn Jahre. Das heisst, dass die Frage in den letzten Jahren kaum Beachtung gefunden hat. Die gängige Praxis, Erhebungen anonym durchzuführen, insbesondere bei psychometrischen Verfahren, wurde daher unbesehen auf Programmevaluationen übertragen. Um den künftigen Evaluationsprozess auf Evidenz abzustützen und um die aktuelle Forschungslücke über anonyme und personalisierte Evaluation zu schliessen, entschieden die Verantwortlichen, eine Untersuchung durchzuführen. Diese sollte folgende Fragestellung bearbeiten: **Unterscheiden sich die Ergebnisse anonymisierter und personalisierter Evaluationen?**

Die Beantwortung dieser Hauptfragestellung erfolgte in drei Teilfragestellungen, die eine differenzierte Analyse hinsichtlich des Untersuchungsziels erlauben:

1. Werden die Module bei der quantitativen Bewertung bei anonym erhobenen Evaluationen im Vergleich zu personalisiert erhobenen Evaluationen besser bewertet?
2. Werden bei anonymen oder bei personalisiert erhobenen Evaluationen häufiger optionale Textantworten gegeben?
3. Unterscheiden sich die Textantworten bei anonymen bzw. personalisierter Evaluation in ihrer Qualität (siehe dazu Kapitel 2.3)?

Methoden

Fragebogen

Die für das Curriculum eigens entwickelten Fragebogen zur Modulevaluation beinhalten sechs Items zum Aspekt **allgemeiner Eindruck** des Moduls (Ziele, Relevanz, Vertiefung, Lernstoff, Organisation, Qualität), vier Items zum Aspekt **Vorlesungen** (Lernfragen, Dozierende, Strukturie-

III Vorlesungen	-- trifft keinesfalls zu			trifft in hohem Masse zu ++		
	--	-	--+	+-	+	++
1. Die Vorlesungen waren geeignet für das Bearbeiten der Lerninhalte						
2. Der Schwierigkeitsgrad der Vorlesungen war angemessen						
3. Die Dozierenden drückten sich gut verständlich aus						
4. Der Lernstoff wurde klar strukturiert präsentiert						

Abbildung 1: Beispiel für die Items zur Bewertung der Vorlesungen

rung, Qualität, siehe Abbildung 1) und vier Items zu den **Prüfungen** (Niveau, Umfang, Inhalt, Qualität). Alle diese Items bewerten die Studierenden auf einer Skala von 1 (trifft keinesfalls zu) bis 6 (trifft in hohem Masse zu). Diese Skalierung wurde gewählt, weil sie dem Schweizer Notensystem entspricht und so für die Studierenden bei der Bewertung wenig kognitiven Aufwand bedeutete und weil damit eine Tendenz Richtung genügend/ungenügend erkennbar wurde. Je nach Aufbau der einzelnen Module wurden zudem verschiedene Items, z. B. zu den Tutoraten, Skillstrainings, Übungen oder Seminaren erfragt, siehe Abbildung 1 [1], [2], [9], [http://www.qualitaet.unibe.ch/unibe/qualitaet/content/e4198/e4436/e4456/MusterfragebogenTheologieVorlesung_ger.pdf].

Für die nachfolgenden Analysen zur Untersuchung der Fragestellung wurde jeweils zu jedem Aspekt über die Mittelwertbildung der dazugehörigen Items eine ebenfalls sechsstufige Skala gebildet.

Neben der Bewertung des Moduls mit der vorgegebenen Skala hatten die Studierenden die Möglichkeit, im Fragebogen zu jedem Aspekt Kommentare und/oder Anregungen abzugeben. Das bedeutet, dass die Studierenden im Fragebogen mehrere Textantworten geben konnten.

Stichprobe und Datenbasis

Für die Untersuchung der Fragestellung wurde das Evaluationssetting verändert, d. h. die Hälfte der Studierenden füllte die Fragebögen weiterhin anonym aus, während die andere Hälfte die Fragebögen ausfüllte, die bereits ihren Namen aufgedruckt trugen. Die Datensammlung für diese Untersuchung dauerte zwei Jahre. Die Stichprobe bestand aus Studierenden von vier Kohorten. Bei jeder Modulevaluation wurden die Studierenden jeweils neu durch Randomisierung entweder der Gruppe „anonym“ oder der Gruppe „personalisiert“ zugeteilt.

Während des Erhebungszeitraums wurden insgesamt 27 Module evaluiert. Von der Untersuchung ausgeschlossen waren alle Praxis- und Kommunikationsmodule sowie die Module des Clinical Assessments, da diese durch andere Fragebogen evaluiert wurden. Die Studienverantwortlichen wählten von den 27 Modulen für die Untersuchung fünf Module aus, die den vielfältigen Bachelorstudiengang Pflege angemessen abbilden: „Pflege studieren (Einführungsmodul)“, „Notfallsituationen handhaben“, „Forschung verstehen“, „Statistisches Basiswissen erwerben“ und „Qualität der Pflege sichern“. Insgesamt flossen da-

mit 615 ausgefüllte Fragebogen, davon 306 in anonymer Form, in die Untersuchung ein.

Die Analysen der Antworten wurden pro Modul einzeln und unabhängig voneinander durchgeführt, da die Gruppen bei jedem Modul neu randomisiert wurden, und es somit wahrscheinlich war, dass der/die gleiche Studierende ein Modul in anonymer und ein anderes Modul in personalisierter Form bewertet hatte.

Datenaufbereitung

Für die ersten beiden Teilfragestellungen wurden die Daten in das Statistikprogramm SPSS eingegeben und die in Kapitel 3 beschriebenen Analysen durchgeführt. Für die dritte Fragestellung bedurfte es einer separaten Aufbereitung der Daten. Dafür wurden alle Kommentare anhand eines durch die Autoren entwickelten Ratinginstruments bewertet. Das Ziel in der Entwicklung des Instruments war es, Aussagen danach zu kodieren, ob sie vernichtend oder verletzend sind. Dazu wurde zunächst überlegt, was den vernichtenden oder verletzenden Kommentaren gemeinsam ist. Ausgehend davon wurde definiert, dass eine Aussage dann verletzend oder vernichtend ist, wenn sie alle folgenden fünf Bedingungen erfüllt: sie ist

1. negativ,
2. absolut,
3. bezieht sich auf eine Person,
4. enthält einen Befehl und
5. ist optisch akzentuiert.

Daraus wurden die folgenden fünf Dimensionen entwickelt:

1. In der Dimension **Wertung** wurden die Aussagen dahingehend kodiert, ob sie positiv, negativ oder beides waren, z. B. „Das Modul war schlecht organisiert, aber die Themen waren alle sehr interessant“.
2. In der Dimension **Differenziertheit** wurde bewertet, ob die Aussagen differenziert oder absolut waren. Eine Aussage galt z. B. dann als differenziert, wenn sie Worte wie teilweise, manchmal, aber etc. enthielten. Aussagen wie: „Thema Normalverteilung ist unwichtig“ wurden als absolut kodiert.
3. Die Dimension **Individuum** zeigte, ob sich die Aussage auf eine Person bzw. eine/n Dozierende/n bezog oder nicht, ob also deren/dessen Name erwähnt wurde, z. B. „Frau Müller ist unfähig!“.

4. In der Dimension **Befehl** wurde beurteilt, ob die Rückmeldungen eine Forderung enthielten, z. B. „Erklären Sie den Lehrbeauftragten endlich, wie der Beamer funktioniert“.
5. In der Dimension **Optische Akzentuierung** wurden Aussagen kodiert, die durch eine Interpunktion wie ein Ausrufezeichen oder eine Unterstreichung hervorgehoben wurden, z. B. „Super!!!“, „HALLO PLANUNG!“.

Alle Kommentare und Anregungen wurden von drei Personen unabhängig voneinander und mit dem vorgestellten Ratinginstrument in kategorialen Variablen kodiert, analysiert und beurteilt. Beim Kodierungsprozess war für die Rater nicht ersichtlich, ob die Aussagen aus der anonymen oder personalisierten Gruppe stammten. Um die Übereinstimmung zwischen den Ratern zu prüfen, wurde der Korrelationskoeffizient nach Pearson gerechnet. Das Maß der Übereinstimmung bei den einzelnen Modulen betrug zwischen $r=.74$ und $r=.98$.

Ergebnisse

Teilfragestellung 1: Unterschiede bei der quantitativen Bewertung

Zur Untersuchung, ob anonym erhobene Evaluationen negativer ausfallen als personalisiert erhobene, wurde für jedes einzelne Modul eine multivariate Varianzanalyse (MANOVA) durchgeführt. Abhängige Variablen (AVs) waren die gebildeten Skalen zu den beschriebenen Aspekten der Modulevaluation, die unabhängige Variable (UV) war die Gruppierungsvariable „anonym“ und „personalisiert“. Bei keinem Modul konnte ein signifikanter Unterschied zwischen anonym und personalisiert erhobenen Daten nachgewiesen werden (siehe Tabelle 1).

Teilfragestellung 2: Häufigkeit der Textantworten

Nach einer einfachen Häufigkeitszählung der optionalen Textantworten wurde pro Modul ein Mann-Whitney-Test für unabhängige Stichproben durchgeführt zur Prüfung, ob die anonym evaluierenden Studierenden häufiger Textantworten geben. Der Mann-Whitney-Test wurde gewählt, da die Daten die statistischen Voraussetzungen für einen T-Test nicht erfüllten.

Wie in Tabelle 2 ersichtlich ist, geben die Studierenden aus der personalisierten Gruppe tendenziell häufiger Anregungen oder Kommentare ab. Die Häufigkeit unterscheidet sich jedoch nicht signifikant zwischen den Gruppen. Als Ausnahme gelten die Textantworten im Modul „Forschung verstehen“, bei dem die personalisierte Gruppe im Durchschnitt signifikant häufiger eine Anregung vermittelt oder einen Kommentar abgegeben hat.

Teilfragestellung 3: Qualität der Textantworten

Insgesamt haben die Studierenden über alle fünf für die Untersuchung relevanten Module $n=2152$ Aussagen formuliert, davon waren total $n=6$ verletzend oder vernichtend, erfüllten also alle fünf genannten Bedingungen (siehe Kapitel 2.3 Datenaufbereitung). Aufgrund dieser geringen Anzahl von Aussagen, die die Kriterien einer verletzenden bzw. vernichtenden Aussage erfüllten, konnte keine statistische Analyse durchgeführt werden, um zu untersuchen, ob verletzende bzw. vernichtende Aussagen häufiger von anonymen Studierenden stammen.

Für die statistische Analyse wurde in einem ersten Schritt pro Dimension ein Summenwert über alle gegebenen Textantworten im Fragebogen gebildet. Das heißt, wenn ein Studierender in einer Evaluation drei Textantworten gab, die alle nicht persönlich waren, erhielt er auf der Dimension Individuum einen Summenwert=3. Je niedriger also der Summenwert war, desto häufiger haben die Studierenden negative, absolute, persönliche, einen Befehl oder optische Akzentuierung enthaltende Rückmeldungen formuliert.

In einem zweiten Schritt wurden unter Anwendung von MANOVAs (UV=Gruppierungsvariable anonym vs. personalisiert, AVs=Summenwerte aller Dimensionen) mögliche Gruppenunterschiede in der Qualität der Textantworten geprüft (siehe Tabelle 3).

Bei keinem Modul konnten zwischen den personalisierten und anonymen Textantworten signifikante qualitative Unterschiede in einer der fünf Dimensionen nachgewiesen werden.

Diskussion

Ziel dieser Studie war es, herauszufinden, ob ein Unterschied zwischen anonymer und personalisierter Programmevaluation durch Studierende besteht. Als Ergebnis stellte sich heraus, dass es hinsichtlich der quantitativen Bewertungen keine Unterschiede zwischen den beiden untersuchten Gruppen bestehen. Wegen der geringen Datenmenge konnte nicht untersucht werden, ob es einen Unterschied hinsichtlich verletzender bzw. vernichtender Aussagen gibt, allerdings lässt sich bei der Untersuchung der Ausprägung der fünf qualitativen Dimensionen kein signifikanter Unterschied zwischen anonymer und personalisierter Evaluation nachweisen. Bei der Häufigkeit der Textantworten ergab sich einzig bei einem Modul ein signifikanter Unterschied, bei allen anderen Modulen konnte kein Unterschied zwischen der anonymen und personalisierten Gruppe nachgewiesen werden. Diese Ergebnisse öffnen einen weiten Interpretationsspielraum. Für das einzige signifikante Ergebnis könnte die Tatsache verantwortlich sein, dass Studierende, die sich die Mühe nehmen, einen reflektierten persönlichen Kommentar zu formulieren, als Personen wahrgenommen werden möchten. Sie würden gerne einen weiterführenden Austausch beginnen.

Tabelle 1: Ergebnisse der MANOVAs pro Modul zur quantitativen Bewertung

Modul	N(anonym)	Pillai-Spur	F-Wert (df)	P-Wert
Pflege studieren	119(61)	0.04	0.83(6)	.547
Notfallsituationen handhaben	93(47)	0.05	0.74(6)	.621
Forschung verstehen	134(66)	0.01	0.43(4)	.785
Statistisches Basiswissen erwerben	159(79)	0.02	0.89(4)	.470
Qualität der Pflege sichern	110(53)	0.02	0.34(6)	.914

Tabelle 2: Häufigkeit der Textantworten

Modul	Gruppe	M(SD)	U-Wert	Z-Wert
Pflege studieren	anonym	3.49(2.16)	3276.50	-1.43
	personalisiert	3.95(1.98)		
Notfallsituationen handhaben	anonym	3.94(1.79)	1045.50	-1.43
	personalisiert	4.52(1.34)		
Forschung verstehen	anonym	1.84(1.40)	2654.50*	-2.26
	personalisiert	2.33(1.34)		
Statistisches Basiswissen erwerben	anonym	1.68(1.45)	2840.00	-1.25
	personalisiert	1.99(1.53)		
Qualität der Pflege sichern	anonym	4.00(1.68)	1567.00	-0.33
	personalisiert	3.88(1.61)		

* p≤.05 (1-seitig)

Tabelle 3: Ergebnisse der MANOVAs pro Modul zu den qualitativen Dimensionen

Modul	N(anonym)	Pillai-Spur	F-Wert (df)	P-Wert
Pflege studieren	119(61)	.05	1.67 (5)	.144
Notfallsituationen handhaben	93(47)	.09	1.82 (5)	.116
Forschung verstehen	134(66)	.05	1.23 (5)	.300
Statistisches Basiswissen erwerben	159(79)	.03	0.61 (5)	.695
Qualität der Pflege sichern	110(53)	.04	0.82 (5)	.541

Der unter 1.1 Ausgangslage vorgestellte konsequente Feedbackzyklus und die hohe Transparenz im Umgang mit den Ergebnissen erzeugen bei den Studierenden eine Sicherheit, dass Rückmeldungen ernst genommen und sorgfältig geprüft werden. Die Studierenden missbrauchen deshalb Fragebogen nicht, um anderweitig begründete Frustrationen zu kompensieren.

Neben der standardisierten Modulevaluation bestehen andere Möglichkeiten für Rückmeldungen, insbesondere ein Mentorkonzept, das einen individuellen Austausch zwischen Studierenden und Dozierenden ermöglicht. Allfällige Unzufriedenheiten bzw. Probleme können somit gezielt und auf unterschiedlichen Wegen thematisiert werden. Eine weitere Begründung für die vorliegenden Ergebnisse dürfte sein, dass die Dozierenden signalisierten, dass Rückmeldungen und Vorschläge zur Entwicklung des Studienprogramms willkommen seien.

Dieses Evaluationskonzept hat Auswirkungen sowohl auf die Dozierenden als auch auf die Studierenden. Beide Seiten sehen sich bis zu einem gewissen Grad weiterhin als Pioniere des neuen Studiums und möchten einen Beitrag zur Curriculumsgestaltung leisten. Es kann zudem davon ausgegangen werden, dass in den Lehrplänen der Zubringerschulen (Gymnasien, Berufs- und Fachmittelschulen) Kommunikations- und Feedbackregeln fest inte-

griert sind. Heute studiert eine Generation, die an Austausch und Diskussion gewöhnt ist.

Vernichtende und verletzende Äusserungen müssen aufgrund der Ergebnisse dieser Untersuchung als Einzelfälle betrachtet werden. In den Anfängen der Anwendung des Evaluationskonzepts (1.1 Ausgangslage) war den Dozierenden diese Tatsache noch nicht bewusst und die damalige Datenmenge erlaubte keine generalisierbaren Rückschlüsse. Ausreisser wurden nicht als solche erkannt und deswegen überbewertet. Es wurde vermutet, dass sie wegen der anonymen Durchführung zustande gekommen seien.

Die in der Literatur als „Negativity Bias“ bezeichnete Theorie bestätigt, dass in den meisten Situationen Negatives stärker, dominanter und einschneidender wahrgenommen wird als Positives [10]. Royman und Rozin [11] untermauern diese These plakativ mit dem Vergleich, dass die kurze Berührung mit einer Kakerlake eine delikate Mahlzeit ungeniessbar mache. Gemäss dem Prinzip der „Negativity Dominance“ [11] sollen das Wahrnehmen und Beurteilen von Geschehnissen, die sowohl positive als auch negative Aspekte aufweisen, negativer sein als die arithmetische Summe dieser subjektiven Werte.

Für die vereinzelt negativen Aussagen mögen zahlreiche andere Gründe verantwortlich sein: fachliche Überforderung, private Probleme, Desinteresse an einem

Thema oder ein „Negativity Bias“ auf Seiten der Studierenden – eine zu schwierige Prüfungsfrage macht das ganze Modul schlecht.

Als Limitation der vorliegenden Studie könnte die Eigenkonstruktion des fünfdimensionalen Instruments zur Erfassung der Aussagequalität der optionalen Textantworten genannt werden, da diese Kategorisierung erstens subjektiv ist und zweitens nicht erschöpfend sein kann. Für die Validität des Instruments spricht allerdings die hohe Beurteilungsübereinstimmung zwischen den unabhängigen Ratern. Zu erwähnen bleibt, dass es gerade die Anwendung dieses Ratinginstruments bzw. die Prüfung der optionalen Textantworten war, die Einblick in die kritische Auseinandersetzung der Studierenden mit dem Studienprogramm verschafft hat.

Eine Stärke der Studie ist die grosse Datenmenge, die dank der lückenlosen Durchführung der Modulevaluationen mit einer Rücklaufquote von nahezu 100 Prozent vorliegt. Dies wiederum war nur möglich durch das seriös entwickelte Konzept. Die Evaluationsfragen haben sich über die Jahre hinweg als relevant erwiesen. All diese sorgfältigen Vorarbeiten sowie das eingespielte Setting haben eine komplikationslose Durchführung des Forschungsprojekts ermöglicht.

Schlussfolgerung

Es lohnt sich, in die Entwicklung eines Evaluationskonzeptes zu investieren, das als zentralen Aspekt einen geschlossenen und transparenten Feedbackzyklus vorsieht. Dieser eröffnet den Studierenden Möglichkeiten der Beteiligung und zeigt ihnen, wie die Daten der Evaluationen in die kontinuierliche Entwicklung des Lehrangebots einfließen. Nicht die Anzahl der durchgeführten Evaluationen, wobei ein Minimum sicherlich gewährleistet sein sollte, sondern die konsequente Rückmeldung der Ergebnisse an die Studierenden und die Diskussion über allfällige Massnahmen sind dabei entscheidend. Die Studie zeigt, dass es nicht darauf ankommt, ob anonym oder unter Angabe des Namens evaluiert wird, solange diese konzeptuellen Evaluationsbedingungen erfüllt sind. In diesem Sinne können diese Studienergebnisse auf andere Institutionen übertragen werden.

Bei extrem negativen Antworten muss immer die Frage der Verhältnismässigkeit gestellt werden, da Negatives offensichtlich eine grössere Wirkung hat als Positives. Dies ist eine weitere, möglicherweise erleichternde Schlussfolgerung.

Anmerkung

¹ Schweizer Notensystem: 6 = sehr gut, 5 = gut, 4 = genügend, 3 = ungenügend, 2 = schlecht, 1 = sehr schlecht

² Name geändert

Interessenkonflikt

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

Literatur

1. Kern DE, Thomas PA, Hughes MT. Curriculum Development for Medical Education - A Six-Step Approach. Baltimore: The Johns Hopkins University Press; 1998.
2. Kromrey H. Evaluation - ein vielschichtiges Konzept: Begriff und Methodik von Evaluierung und Evaluationsforschung. Sozialwiss Berufspraxis. 2001;24(2):105-131.
3. Borg WR, Gall MD. Educational research: An introduction. New York: Longman; 1983.
4. Braskamp LA, Ory JC. Assessing faculty work: Enhancing individual and institutional performance. San Francisco: Jossey-Bass; 1994.
5. Seldin P. How administrators can improve teaching: Moving from talk to action in higher education. San Francisco: Jossey-Bass; 1990.
6. Fries CJ, McNinch RJ. Signed versus unsigned student evaluations of teaching: A comparison. Teach Sociol. 2003;31(3):333-344. DOI: 10.2307/3211331
7. Oprea C. The susceptibility of student evaluation of lecturers to situational variables. High Educ. 1980;9(3):293-306. DOI: 10.1007/BF00138519
8. Goh JW, Lee OK, Salleh H. Self-rating and respondent anonymity. Educ Res. 2010;52(3):229-245. DOI: 10.1080/00131881.2010.504060
9. Baartmans P. Qualität nach Mass: Entwicklung und Implementierung von Qualitätsverbesserungen im Gesundheitswesen. 2. ed. Bern: Huber Verlag; 2006.
10. Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD. Bad is Stronger Than Good. Rev Gen Psychol. 2001;5(4):323-370. DOI: 10.1037/1089-2680.5.4.323
11. Rozin P, Royzman EB. Negativity Bias, Negativity Dominance, and Contagion. Personality and Social Psychology Review. 2001;5(4):296-320. DOI: 10.1207/S15327957PSPR0504_2

Korrespondenzadresse:

Theresa Scherer
 Berner Fachhochschule, Studiengang Pflege,
 Murtenstrasse 10, CH-3008 Bern, Schweiz
 theresa.scherer@bfh.ch

Bitte zitieren als

Scherer T, Straub J, Schnyder D, Schaffner N. Der Einfluss von Anonymität in der Lehrevaluation durch Studierende. GMS Z Med Ausbild. 2013;30(3):Doc32.
 DOI: 10.3205/zma000875, URN: urn:nbn:de:0183-zma0008755

Artikel online frei zugänglich unter

<http://www.egms.de/en/journals/zma/2013-30/zma000875.shtml>

Eingereicht: 12.11.2012

Überarbeitet: 31.01.2013

Angenommen: 07.04.2013

Veröffentlicht: 15.08.2013

Copyright

©2013 Scherer et al. Dieser Artikel ist ein Open Access-Artikel und steht unter den Creative Commons Lizenzbedingungen (<http://creativecommons.org/licenses/by-nc-nd/3.0/deed.de>). Er darf vervielfältigt, verbreitet und öffentlich zugänglich gemacht werden, vorausgesetzt dass Autor und Quelle genannt werden.

The Effects of Anonymity on Student Ratings of Teaching and Course Quality in a Bachelor Degree Programme

Abstract

Research Question: Are there any clear differences between the outcomes of anonymous and personalised student evaluations of teaching quality?

Methods: During a two-year period students were randomly divided into two separate groups, “anonymous” and “personalised”, for end-of-module evaluations. The quality of the module was assessed using a standardised questionnaire. Additionally, students were given the option to add “further comments” if they wanted to highlight specifics.

These optional comments were independently assessed by three people, using a five-dimensional rating instrument: positive/negative; differentiated/absolute; naming a person/general; containing an order/neutral; visually accentuated/blank.

The database consisted of 615 evaluation forms, of which 306 were completed anonymously. In order to identify whether there were any differences between the anonymous and personalised data, a multivariate variance analysis was performed.

Based on the scale, the answers to the questions and the quality of the comments were evaluated. Furthermore, an assessment was made to determine if there were any differences in the number of optional comments between the two groups.

Results: No significant differences were identified in the informative quality of data between the anonymous and personalised student evaluations. However, students in the personalised group had a tendency to include more details in their written answers.

Conclusion: Personalised evaluations do not generate more biased results in terms of social desirability, as long as the evaluation concept is characterised by a closed-circle process and is transparent. In other words, it is imperative that the outcomes of the evaluation are reported back to the students. Moreover, there has to be an opportunity for students to discuss any further suggestions and/or future desires in an open environment. In this way the students respect and understand that their feedback is being taken seriously; consequently, they feel able to provide a constructive and honest evaluation.

Keywords: Education, curriculum development, programme evaluation, respondent anonymity

Theresa Scherer¹

Jan Straub¹

Daniel Schnyder¹

Noemi Schaffner¹

1. Bern University of Applied Sciences, degree programme Nursing, Bern, Schweiz

Introduction

Initial Situation

In 2006, the Bachelor of Science in Nursing at the Bern University of Applied Sciences (BFH) was redesigned in accordance with the requirements of the Bologna Process. The degree programme is a competence-based general studies course, combining a scientific basis and consistent practical focus. The structure is modular; the pedagogic-didactic concept is Problem-based Learning (PBL). The teaching staff consists of twenty lecturers who have

generally completed basic training in nursing and hold an academic degree in Nursing Science, Educational Science or Psychology. On average, 100 students are educated in Nursing Science every academic year.

Based on the evaluation literature which has proven itself well for PBL curriculums, the evaluation process of the degree programme was developed in parallel to the curriculum [1], [2]. The objective of the teaching evaluation (evaluations of modules, lecturers, etc.) was the continuous optimisation of the degree programme. This led to a corresponding evaluation concept, which can be characterised as follows:

1. **Continuous evaluation of all modules by students and lecturers** for structural-organisational adaptations within the single modules.
2. **Continuous curriculum adaptation by the study group.** This is the evaluation group, consisting of the head of the degree programme, the head of the training programme and the research associate. It sees to structural, content-orientated adaptations within the whole curriculum.
3. **Continuous quality development by the quality circle.** The quality circle is composed of the entire teaching staff. It sees to the systemic-attitudinal adaptations, for example, concerning the basic educational attitude, for the realisation of the future professional profile.

At the end of every module, the students evaluated the module using a standardised questionnaire. The questions were designed to monitor the module quality. The students completed the questionnaire during a teaching session at the end of every module. This resulted in the very high return rate of 95 per cent. All evaluations were carried out anonymously.

The collected data was processed statistically and summarised. In the following module, the head of the degree programme presented the summary to the students and the results were discussed. The students received feedback as to whether and how their proposed changes or suggestions would be put into practice.

This approach has proven itself well in the past years. However, the question of whether the student survey should still be carried out anonymously led to controversial discussions among the staff. The reason for this was the appearance of single, noticeably negative numerical assessments, for example, rating all the questions with 1¹ or giving offensive and destructive comments such as “Mrs Müller² is completely incompetent! Send her to a further education course!” In addition, the very negative comments from the students were often not consistent with the corresponding numerical assessments. In such cases, it was not clear to the teaching staff how the results should be interpreted. Since the students evaluated the modules without giving their names, the lecturers could not check back with them, which led to great irritation. Part of the teaching staff suspected that anonymous evaluation favoured such very negative student responses. However, the amount of data available at the time was very small and therefore no well-founded conclusions could be drawn.

Supporters of the anonymous approach argued that students might fear that negative criticism could result in unpleasant consequences from lecturers. The research literature also points out that, particularly with sensitive or threatening answers that could have negative effects, there is a risk that the questionnaire will not be completed honestly but rather in accordance with social desirability [3]. Moreover, various studies have provided evidence that respondents tend to give more obliging answers if they know or suppose that the lecturers will be able to

see their names alongside their answers [4], [5]. According to another study, carried out by Fries and McNinch [6], students do not give their names, even when asked to do so, whenever they have something negative to say. Therefore the specialised literature recommends that questionnaires should be completed anonymously in order to ensure the accuracy and reliability of the data [3]. However, there have been just as many studies that could not prove that there are any differences between anonymous and personalised data [7], [8]. Furthermore, the teaching staff adopts an attitude, in accordance with the basic educational concept, that students should be treated as adult partners and that an objective and open exchange regarding the performance quality must be possible. In the course of their studies, the students should learn to be able to give critical yet respectful and responsible feedback.

Research Question

Previous research results on the subject of anonymous and personalised evaluation are inconsistent and some of the specialised literature is more than ten years old. This means that the question has not received much attention in recent years. The common approach to carrying out surveys anonymously – particularly when using psychometric methods – was therefore applied indiscriminately to programme evaluations. To be able to base the evaluation process on evidence and to close the current research gap regarding anonymous and personalised evaluation, those responsible decided to carry out a study. Its aim was to discuss the following question: **Are there any differences between the outcomes of anonymous and personalised evaluations?**

To answer this core research question, three sub-questions were defined, allowing a differentiated analysis regarding the research objective:

1. Do the modules receive better quantitative assessment if the evaluations are anonymous or if they are personalised?
2. Are more optional comments given in anonymous or personalised evaluation?
3. Is the quality of the optional comments different in anonymous and personalised evaluation (see Chapter 2.3 “Data Processing”)?

Methods

Questionnaire

The module evaluation questionnaires, specially designed for the curriculum, contain six items for the aspect **Overall Impression** of the module (objectives, relevance, specialism, subject matter, organisation, quality), four items for the aspect **Lectures** (didactic questions, lecturers, structure, quality, see Figure 1) and four items for the aspect **Exams** (level, scale, content, quality). The students assessed all of these items on a scale from 1 (completely

III Lectures	-- completely disagree			completely agree ++		
	--	-	-+	+ -	+	++
1. The lectures were suitable to deal with the subject matter						
2. The level of difficulty of the lectures was appropriate						
3. The lecturers expressed themselves comprehensibly						
4. The subject matter was presented in a clearly structured way						

Figure 1: Example of items for the assessment of lectures

disagree) to 6 (completely agree). This scaling was chosen because it is consistent with the Swiss grading system and therefore demands little cognitive effort from the students while evaluating and because it allowed a tendency towards sufficient/insufficient to be observed. In addition, depending on the structure of the individual modules, different items were questioned, e.g. concerning tutorials, skill trainings, practical courses or seminars, see Figure 1 [1], [2], [9], [http://www.qualitaet.unibe.ch/unibe/qualitaet/content/e4198/e4436/e4456/MusterfragebogenTheologieVorlesung_ger.pdf].

To be able to subsequently analyse the research question, an additional six-level scale was designed for every aspect by averaging its items. In addition to the module assessment using the given scale, the students had the possibility to make comments and/or suggestions about every aspect. This means that the students could make more than one optional comment in the questionnaire.

Sample and Database

For the analysis of the research question the evaluation setting was changed: one half of the students still completed the questionnaires anonymously, whereas the other half completed questionnaires with their names printed on them. The data collection for this study took two years. The sample was composed of students from four cohorts. For every module evaluation the students were randomly re-assigned either to the “anonymous” group or the “personalised” group.

During the study period a total of 27 modules were evaluated. All the practice and communication modules as well as the Clinical Assessment modules were excluded from the study because other questionnaires were used to evaluate these. The people in charge of the study chose five out of the 27 evaluated modules that suitably represent the multifaceted Bachelor in Nursing: “The study of nursing (introductory module)”, “Dealing with emergency situations”, “Understanding research”, “Acquiring basic statistical knowledge” and “Ensuring nursing quality”. Altogether 615 completed questionnaires, of which 306 were completed anonymously, were used in the study.

The analyses of the answers were carried out individually and independently for each module. Since the groups for every module evaluation had been randomised anew, it was quite likely that the same student assessed one

module anonymously and another module with a personalised questionnaire.

Data Processing

For the first two sub-questions, the data was entered into the statistical programme SPSS and the analysis carried out as described in Chapter 3 “Results”.

For the third sub-question the data had to be processed separately. In order to do this, all the comments were assessed using a rating instrument developed by one of the authors. The instrument was developed with the objective of coding statements in a way that shows whether they are offensive or destructive. First, the common features of offensive or destructive comments had to be considered. Based on this, it was defined that a statement is offensive or destructive only if it meets all of the following five conditions: it is

1. negative,
2. absolute,
3. refers to a person,
4. contains an order and
5. is visually accentuated.

The five resulting dimensions were defined as follows:

1. In the dimension **Assessment** the statements were coded as positive, negative or both, e.g. “The module was poorly organised, but the topics were all very interesting”.
2. In the dimension **Differentiation** it was assessed whether the statements were differentiated or absolute. A statement, for example, was considered differentiated if it contained words like “partly”, “sometimes”, “but”, etc. Statements like: “The topic of normal distribution is unimportant” were coded as absolute.
3. The dimension **Individual** showed whether a statement referred to a person, namely a lecturer, and whether her/his name was mentioned, e.g. “Mrs Müller is incompetent!”
4. In the dimension **Order** it was assessed whether the responses contained an order, e.g. “Explain to the associate lecturers once and for all how the digital projector works”.
5. In the dimension **Visual Accentuation**, statements were coded as to whether they were accentuated with

punctuation like an exclamation mark or highlighting, e.g. "Super!!!" or "HELLO PLANNING!".

Using the rating instrument presented above, three people independently coded all comments and suggestions into categorical variables and analysed and assessed them. During the coding process the assessors did not know whether a statement came from the anonymous or the personalised group. To verify the assessors' consistency, the Pearson correlation coefficient was calculated. The value of the consistency for the single modules was between $r=.74$ and $r=.98$.

Results

Sub-Question 1: Differences in the Quantitative Assessment

To analyse whether anonymous evaluations tended to be more negative than personalised evaluations, a multivariate variance analysis (MANOVA) was carried out for every single module. The scales developed according to the described aspects of the module evaluation were the dependent variables; the grouping variable "anonymous" and "personalised" was the independent variable.

For none of the modules could any significant difference between anonymous and personalised data be identified (see Table 1).

Sub-Question 2: Frequency of Optional Comments

After a simple frequency count of the optional comments, a Mann-Whitney test for independent samples was carried out for every module to verify whether the students evaluating anonymously tended to make optional comments more frequently. The Mann-Whitney test was chosen because the data did not meet the requirements for a t-test. As can be seen in Table 2, the students in the personalised group tended to make suggestions and comments more often. However, the frequency does not differ significantly from one group to the other. The optional comments in the module "Understanding research" can be considered as an exception: in this module, the personalised group made suggestions or expressed opinions significantly more often on average.

Sub-Question 3: Quality of Optional Comments

The students formulated a total of $n=2152$ statements about the five modules relevant for the study; $n=6$ of these were offensive or destructive, meaning they met all of the above-mentioned conditions (see Chapter 2.3 "Data Processing"). Due to this small number of statements meeting the conditions of an offensive or destructive statement, it was not possible to carry out a statistical analysis to determine whether the students evaluating

anonymously made offensive or destructive statements more often.

As a first step for the statistical analysis, the cumulative value per dimension was established for all the optional comments in the questionnaire. This means that if a student made three non-personal comments in one evaluation, he or she received a cumulative value of 3 on the dimension "Individual". Thus the lower the cumulative value, the more often the students wrote responses that were negative, absolute or personal or that contained an order or a visual accentuation.

As a second step, the data was checked for possible differences between the groups in the quality of the optional comments using MANOVA (IV=grouping variable anonymous vs. personalised, DVs=cumulative values of all dimensions) (see Table 3).

No significant qualitative differences between anonymous and personalised comments could be identified in any of the five dimensions.

Discussion

The objective of the study was to determine whether there is a difference between anonymous and personalised programme evaluation by students. As a result it could be shown that, regarding the quantitative assessments, there are no differences between the two analysed groups. Due to the small amount of data it could not be determined whether there is a difference concerning offensive or destructive statements, but an analysis of the characteristics of the five qualitative dimensions shows no significant difference between anonymous and personalised evaluation. The frequency of the optional comments differed significantly only for one module; for all the other modules no difference between the anonymous and personalised group could be found. These results leave much room for interpretation. The single significant result might be explained by the fact that students who make an effort to formulate an elaborate personal comment would like to be perceived as individuals. They would like to initiate a further exchange.

The consistent feedback cycle presented in Chapter 1.1 "Initial Situation" and the high degree of transparency in the handling of the results gives the students the certainty that their feedback will be taken seriously and considered carefully. The students therefore do not abuse questionnaires to vent frustration about other issues.

In addition to the standardised module evaluation other feedback possibilities exist, notably a mentoring concept that facilitates individual exchange between students and lecturers. Possible dissatisfactions or problems can thus be addressed specifically and in various ways. Another explanation for the present results might be that the lecturers signalled that feedback and suggestions for the development of the programme would be welcome. This evaluation concept affects lecturers as well as students. Both still see themselves, to a certain extent, as pioneers of the new degree programme and would like

Table 1: Results of the MANOVA for each module with quantitative assessment

Module	N(anonymous)	Pillai's trace	F-value (df)	p
The study of nursing	119(61)	0.04	0.83(6,112)	.547
Dealing with emergency situations	93(47)	0.05	0.74 (6,86)	.621
Understanding research	134(66)	0.01	0.43(4,129)	.785
Acquiring basic statistical knowledge	159(79)	0.02	0.89(4,154)	.470
Ensuring nursing quality	110(53)	0.02	0.34(6,103)	.914

Table 2: Frequency of comments

Module	Group	M(SD)	U-Value	Z-Value
The study of nursing	anonymous	3.49(2.16)	3276.50	-1.43
	personalised	3.95(1.98)		
Dealing with emergency situations	anonymous	3.94(1.79)	1045.50	-1.43
	personalised	4.52(1.34)		
Understanding research	anonymous	1.84(1.40)	2654.50*	-2.26
	personalised	2.33(1.34)		
Acquiring basic statistical knowledge	anonymous	1.68(1.45)	2840.00	-1.25
	personalised	1.99(1.53)		
Ensuring nursing quality	anonymous	4.00(1.68)	1567.00	-0.33
	personalised	3.88(1.61)		

* p \leq .05 (1-sided)**Table 3: Results of the MANOVA for each module for the qualitative dimensions**

Module	N(anonymous)	Pillai's trace	F-value (df)	p
The study of nursing	119(61)	.05	1.67 (5,149)	.144
Dealing with emergency situations	93(47)	.09	1.82 (5,91)	.116
Understanding research	134(66)	.05	1.23 (5,126)	.300
Acquiring basic statistical knowledge	159(79)	.03	0.61 (5,117)	.695
Ensuring nursing quality	110(53)	.04	0.82 (5,107)	.541

to contribute to the curriculum design. Furthermore it can be assumed that rules for communication and feedback are fully integrated in the curricula of the feeder schools, where the higher education entrance qualification is obtained. The generation studying today is used to communication and discussion.

Based on the results of the study, offensive and destructive statements must be viewed as isolated cases. The lecturers were unaware of this fact when the evaluation concept was first used (see Chapter 1.1 "Initial Situation"), and the amount of data available at the time did not allow any universally applicable conclusions to be drawn. Outliers were not recognised as such and therefore overrated. It was suspected that they came about because of the use of anonymity.

The theory referred to as "Negativity Bias" in the literature confirms that in most situations something negative will be perceived more strongly, dominantly and drastically than something positive [10]. Royman and Rozin [11] illustrate this theory with the comparison that the slightest

contact with a cockroach makes a delicious meal inedible. According to the principle of "Negativity Dominance" [11], the perception and judgement of events that feature both positive and negative aspects are more negative than the arithmetical sum of these subjective values.

Numerous other reasons may be accountable for single negative statements: professional overload, private problems, lack of interest in a subject or a "Negativity Bias" on the students' side, for example, one exam question that is too difficult makes the whole module bad.

As a limitation of the present study one could name the self-elaborated design of the five-dimensional instrument to determine the statement quality of the optional comments, since this categorisation is subjective and cannot be exhaustive. Nevertheless, the high assessment consistency between the independent assessors is indicative of the validity of the instrument. It should also be noted that it is precisely the use of this rating instrument and the assessment of the optional comments that made it

possible to gain an insight into the students' critical engagement with the curriculum.

A strength of the study lies in the large amount of data available: due to the comprehensive implementation of the module evaluation, the response rate was almost 100 per cent. This in turn was only made possible by the well-thought-out concept. Over the years, the evaluation questions have proven to be relevant. All the careful preliminary work and the well-practised setting made a smooth implementation of the research project possible.

Conclusion

It is worthwhile to invest in the development of an evaluation concept whose central aspect is a closed and transparent feedback cycle. This provides the students with the possibility of participation and shows them how the evaluated data is integrated into the continuous development of the course offer. What is crucial is not the number of evaluations – although a minimum should certainly be ensured – but rather the consistent reporting of the results to the students and the discussing of possible measures. The study shows that it does not matter whether an evaluation is carried out anonymously or in a personalised way, as long as these conceptual evaluation conditions are met. With this in mind, the results of the present study can be applied to other academic institutions.

Extremely negative responses must be put into perspective, since negativity seems to have a stronger effect than positivity. This is a further, possibly relieving, conclusion.

Note

¹ Swiss grading system: 6 = excellent, 5 = good, 4 = pass, 3 = fail, 2 = very poor, 1 = no performance

² Real name withheld.

Competing interests

The authors declare that they have no competing interests.

References

1. Kern DE, Thomas PA, Hughes MT. Curriculum Development for Medical Education - A Six-Step Approach. Baltimore: The Johns Hopkins University Press; 1998.
2. Kromrey H. Evaluation - ein vielschichtiges Konzept: Begriff und Methodik von Evaluierung und Evaluationsforschung. Sozialwiss Berufspraxis. 2001;24(2):105-131.

3. Borg WR, Gall MD. Educational research: An introduction. New York: Longman; 1983.
4. Braskamp LA, Ory JC. Assessing faculty work: Enhancing individual and institutional performance. San Francisco: Jossey-Bass; 1994.
5. Seldin P. How administrators can improve teaching: Moving from talk to action in higher education. San Francisco: Jossey-Bass; 1990.
6. Fries CJ, McNinch RJ. Signed versus unsigned student evaluations of teaching: A comparison. Teach Sociol. 2003;31(3):333-344. DOI: 10.2307/3211331
7. Oprea C. The susceptibility of student evaluation of lecturers to situational variables. High Educ. 1980;9(3):293-306. DOI: 10.1007/BF00138519
8. Goh JW, Lee OK, Salleh H. Self-rating and respondent anonymity. Educ Res. 2010;52(3):229-245. DOI: 10.1080/00131881.2010.504060
9. Baartmans P. Qualität nach Mass: Entwicklung und Implementierung von Qualitätsverbesserungen im Gesundheitswesen. 2 ed. Bern: Huber Verlag; 2006.
10. Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD. Bad is Stronger Than Good. Rev Gen Psychol. 2001;5(4):323-370. DOI: 10.1037/1089-2680.5.4.323
11. Rozin P, Royzman EB. Negativity Bias, Negativity Dominance, and Contagion. Personality and Social Psychology Review. 2001;5(4):296-320. DOI: 10.1207/S15327957PSPR0504_2

Corresponding author:

Theresa Scherer

Bern University of Applied Sciences, degree programme Nursing, Murtenstraße 10, CH-3008 Bern, Schweiz
theresa.scherer@bfh.ch

Please cite as

Scherer T, Straub J, Schnyder D, Schaffner N. Der Einfluss von Anonymität in der Lehrevaluation durch Studierende. GMS Z Med Ausbild. 2013;30(3):Doc32.

DOI: 10.3205/zma000875, URN: urn:nbn:de:0183-zma0008755

This article is freely available from

<http://www.egms.de/en/journals/zma/2013-30/zma000875.shtml>

Received: 2012-11-12

Revised: 2013-01-31

Accepted: 2013-04-07

Published: 2013-08-15

Copyright

©2013 Scherer et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en>). You are free to Share – to copy, distribute and transmit the work, provided the original author and source are credited.