

Autor: Bernhard Jacobs, Fachrichtung Erziehungswissenschaften der Universität des Saarlandes
10. 7. 2009

Leistungssteigerung durch Notendruck ? - Die Wirkung der Benotung auf die Studierleistungen in einem Seminar

Abstract

Ziel des Feldexperimentes war es, die Auswirkung der Benotung eines Quiz auf die Leistung in diesem Quiz sowie im Hinblick auf das langfristige Behalten zu überprüfen und durch die Erfassung motivationaler Variablen eine Erklärung gemäß der Erwartungs- mal Werttheorie für die erwarteten Befunde zu finden. Studierende aus 4 Seminaren wurden nach Zufall auf 2 Gruppen aufgeteilt und absolvierten dasselbe Quiz entweder unter der Erwartung, das Quiz werde benotet oder nicht. Das verpflichtende, unbenotete Quiz diente vornehmlich Übungszwecken und gewährte vielfältige Item spezifische Rückmeldungen. Im Verlauf des Seminars wurden die Quizbedingungen für beide Gruppen getauscht. Die Ergebnisse sprechen konsistent für deutlich bessere Leistungen in den benoteten Quiz. Der Leistungsvorteil bleibt eine Woche voll erhalten und lässt sich auch noch nach 6 bis 8 Wochen statistisch sichern. Benotete Quiz lösen im Vergleich zu unbenoteten Quiz durchgängig mehr aktuelle Angst aus. Die besseren Quizleistungen unter der Benotungsbedingung lassen sich recht gut mit der deutlich intensiveren Vorbereitung erklären, die ihrerseits durch die eindeutig höhere Bedeutsamkeit, im Quiz gute Leistungen erzielen zu wollen, motiviert wird. Weitere Befunde deuten darauf hin, dass verpflichtende, aber unbenotete Onlinequiz mit vielfältigen Feedbackmöglichkeiten im Vergleich zu keiner Maßnahme einen relativ schwachen, aber teilweise noch signifikant nachweisbaren, Lerngewinn erbringen können.

Schlagworte: extrinsische Motivierung, extrinsische Motivation
Benotungssystem, Notenanreiz, Testen, Prüfung
extrinsic motivation, grading, testing

Problemstellung und Zielsetzung der Untersuchung

Einer trivialen Binsenweisheit zufolge mobilisiert die Androhung einer Benotung die Lernvorbereitung auf einen Test. Benotete Tests (EG) bringen so bessere unmittelbare Testleistungen hervor als unbenotete Tests (KG). Die wenigen mir bekannten empirischen Untersuchungen zu dieser Thematik bestätigen diese Erkenntnis weitgehend, auch wenn das methodische Niveau der Studien nicht immer überzeugen kann (siehe Jacobs 2007). Eine anspruchsvollere Prüfung verlangt den Nachweis, die Testbearbeitung unter benoteten und unbenoteten Quizbedingungen beeinflusse nicht nur die aktuelle Leistung im Quiz, sondern auch in einem späteren Kriteriumstest zugunsten eines stabileren Leistungsvorteils der EG gegenüber der KG. (z.B.: Elikai & Baker 1988). Die Messung selbst soll dabei zumindest nahen Transfer erfassen. Der Anspruch lässt sich beliebig erweitern, indem man dann auch noch Unterschiede in einem viel späteren Abschlusstest zu hoffen glaubt. Schließlich schwebt die These im Raum, durch den verstärkten Einsatz benoteter Tests im Schulsystem ließe sich sogar die Kompetenz der Lerner wirksam steigern. (Bangert-Drowns, Kulik & Kulik (1991).

Ziel der nachfolgenden Untersuchung war es, ein benotetes gegen ein unbenotetes Quiz im Hinblick auf die erzielte Leistung in diesem Quiz zu vergleichen. Darüber hinaus interessierte die Nachhaltigkeit potenzieller Leistungsunterschiede zwischen den Bedingungen. Schließlich sollten auch Daten erhoben werden, um die Motivierung durch eine Benotung näher bestätigen zu können.

Theoretische Annahmen zur Wirkungsweise benoteter Quiz auf die Lernleistung.

Der Studierende kann den potenziellen Lernerfolg in einem Quiz - hier stets verstanden als Test zur Überprüfung der Seminarlehrziele - am ehesten durch eine intensive, angemessene Erarbeitung des prüfungsrelevanten Lehrstoffs beeinflussen. Dazu gehören in erster Linie eine akzeptable Aufmerksamkeit bzw. geistige Mitarbeit im Seminar sowie ein konzentriertes und ausdauerndes Studieren der angebotenen Lernmaterialien mit dem Bestreben um ein hinreichendes Verständnis. Die Bereitschaft, den entsprechenden Lerneinsatz konsequent zu erbringen, wird in hohem Maße von der persönlichen Bedeutung des Quizergebnisses abhängen, die sich durch eine Benotungsandrohung einfach und sehr wirksam erhöhen lässt. Welche Möglichkeiten bleiben, die Lernmotivation auf sonstige Weise anzuregen? Beim verpflichtenden, aber unbenoteten Quiz wurde angekündigt, man erhalte unmittelbare Rückmeldung zu jeder Aufgabe sowie zum Gesamtergebnis und habe so die Möglichkeit, sich selbst zu überprüfen und seine Fehler zu korrigieren. Die Quizbearbeitung könne nützlich im Hinblick auf Abschlussklausur der Vorlesung sein, die für alle Studierenden am Ende des Semesters anstand.

Höchstwahrscheinlich sind die mit der Benotung verbundenen Konsequenzen in ihrer Lernmotivierung zur Quizvorbereitung stärker einzuschätzen als die durch ein unbenotetes Quiz aktivierten Lernanregungen, so dass benotete Quiz bessere aktuelle Leistungsergebnisse nach sich ziehen.

Die experimentelle Untersuchung

Groborientierung zum experimentellen Vorgehen

Im WS 2008 leitete der Verfasser dieser Arbeit 4 Seminare für Studierende des Lehramts im Themengebiet Pädagogische Diagnostik, zwei Seminare mit dem Themenschwerpunkt "Aufgaben stellen und Feedback geben" - im folgenden Seminare Diagnostik Aufgaben genannt - und zwei Seminare mit dem Themenschwerpunkt "Diagnostische Grundlagen" - im folgenden Seminare Diagnostik Grundlagen genannt. 73% aller Seminarteilnehmer waren Frauen. Das Durchschnittsalter betrug 21.5 Jahre. Die Seminare waren so geplant, dass die Untersuchung zugleich wesentliche Übungsbestandteile der Seminare abdeckte. Aus diesem Grunde wurden die Studierenden zu Beginn nicht davon informiert, dass sie zugleich an einem Experiment teilnahmen. Allen Studierenden wurde am Anfang der Veranstaltung mitgeteilt, für den qualifizierten Leistungsnachweis der Seminarendnote sei es notwendig, etliche Quiz zu bearbeiten, die zum Teil benotet und zum Teil unbenotet zu absolvieren seien. Zu Beginn des Seminars konnten die Studierenden im Seminarplan einsehen, zu welchen Terminen und welchen Themen ihr benotetes bzw. unbenotetes Quiz anstand. Die eine Hälfte jedes Seminars bearbeitete dann das benotete Quiz in der Papier -und Bleistift-Version in der entsprechenden Seminarsitzung. Die andere Hälfte des jeweiligen Seminars beantwortete dasselbe Quiz um ein bis 2 Tage verzögert als Computerversion zu Hause via Internet. Um die Behaltenswirkung der Quizbedingung beim ersten Quiz zu überprüfen, fand in der nächsten Seminarsitzung für alle Teilnehmer völlig überraschend eine Leistungsüberprüfung mit einem Parallelquiz statt. Ein Teil der Lehrziele wurde später noch in einer nicht angekündigten unbenoteten Abschlussklausur überprüft.

Unabhängige Variable: Quizbedingung

benotetes Quiz: Die Studierenden bearbeiteten das Quiz im Seminar unter strenger Aufsicht. Nach Abgabe des Quiz zeigte der Seminarleiter unter Zuhilfenahme der Online-Versionen die korrekten Antworten für alle Fragen und gab einige wenige Erklärungen. Die Seminarendnote setzte sich aus den Leistungen der benoteten Quiz, eines benoteten Gruppenarbeitsauftrages und einem Kurzreferat zusammen. Bei den Seminaren Diagnostik Aufgaben lag der Anteil des einzigen benoteten Quiz bei einem Drittel, bei den Seminaren Diagnostik Grundlagen lag der Anteil eines von zwei benoteten Quiz bei einem Viertel der Seminarendnote. Die Bedeutung des Quizergebnisses dürfte grob geschätzt etwas geringer als diejenige einer Klassenarbeit sein.

unbenotetes Quiz: Der Seminarleiter machte zu Beginn des Seminars ganz klar, es sei eine Selbstverständlichkeit, auch das unbenotete Quiz im konstruktiven Sinne zu bearbeiten. Damit nur die berechtigten Studierenden das Onlinequiz überhaupt in Angriff nehmen konnten, wurde jeder berechnigte Student zuvor angemahlt und diesem die URL-Adresse sowie das Passwort mitgeteilt. Zu Beginn des unbenoteten Computerquiz standen klare Anweisungen, das Quiz prüfe wichtige Lehrziele des Seminars, das Ergebnis des Quiz spiele aber keine Rolle für die Seminarendnote. Trotzdem sei man verpflichtet, die Fragen ernsthaft zu beantworten, da offensichtlicher Missbrauch vom System erkannt werde. Am Anfang des Computerquiz wurde die Übung als positive Chance zur Stabilisierung und Verbesserung des Lernens angepriesen sowie etliche konstruktive Ratschläge unterbreitet, wie man das Quiz möglichst optimal bearbeiten sollte. Das Onlinequiz war so konzipiert worden, dass es erst nach Bearbeitung aller Aufgaben abgegeben werden konnte. D.h. Der Studierende hatte theoretisch keine Möglichkeit, irgendeine Aufgabe zu ignorieren. Nach jeder Aufgabenbearbeitung erhielt der Studierende Rückmeldung über die Korrektheit seiner Antwort (knowledge of response (KOR)), die korrekte Antwort (knowledge of correct response (KCR)) sowie in etlichen Fällen elaboriertes Feedback (EL), insgesamt also ein umfangreiches feedback after item, das im Übrigen beliebig oft angewählt werden konnte. Das Computerquiz unterlag keinerlei Zeitbeschränkung. Nach Quizabgabe und

der anonymen Beantwortung einiger Fragen teilte das Programm den Prozentsatz der korrekten Lösungen mit (Knowledge of performance), bestätigte das Ende des Quiz und speicherte die Daten ab. Da die Wirkungsweise von Quiz von vielen Faktoren abhängt, werden in Tabelle 1 zur besseren Übersicht die meiner Meinung nach theoretisch wichtigsten Voraussetzungen und Bedingungen in Form einer Checkliste kurz zusammengefasst.

Tabelle 1: Checkliste der experimentellen Bedingungen

für alle Bedingungen gilt:

- Quizrelevante Seminarsitzung verpflichtend
- Quiztermin angekündigt und lange bekannt
- Quizteilnahme: verpflichtend
- Quizfunktion: Überprüfung zuvor präsentierten Lehrstoffs; Zeit für Vorbereitung ca. 7 Tage.
- Quizlehrziele: sehr transparent, überwiegend Verständnis und Anwendung
- Umfang der Quizlehrziele ca. 1 bis 2 Seminarsitzungen
- angekündigt, dass **kein** benotetes Abschlussexamen (bzw. eine Abschlussklausur) stattfindet.

	EG benotet	KG unbenotet
Quizgewichtung für Seminarnote	25-33 %	0 %
Prüfungsform	streng kontrolliert im Seminar	online zu Hause; ohne externe Kontrolle
Aufgabenbearbeitung	frei, Beantwortung aller Aufgaben wurde angeregt.	alle Aufgaben waren zwingend zu beantworten
Testzeit	relativ großzügig bemessen	beliebig lange
Feedback	KCR after test im Seminar % korrekter Antworten nach einigen Wochen	KOR,KCR und EL after item; % korrekter Antworten unmittelbar nach Quiz

Abhängige Variablen

Als abhängige Leistungsvariablen fungieren die Ergebnisse in den Quiz, im Kriteriumstest, der die Auswirkungen des ersten Quiz überprüfen sollte, sowie in den auf die Quiz abgestimmten Aufgaben der unbenoteten Abschlussklausur. Die einzelnen Aufgaben eines Quiz wurden je nach Testumfang unterschiedlich stark gewichtet. Sofern Aufgaben mehrere Antworten verlangten, wurden sie einer Zufallskorrektur unterzogen. Das jeweilige Quizergebnis ergibt sich aus der Summe der gewichteten Aufgabenpunkte und wird hier relativiert an der Höchstpunktzahl als Prozentsatz der korrekten Antworten mitgeteilt.

Prozentsatz der korrekten Antworten =

$$\left(\frac{\text{Summe der erzielten Aufgabenpunkte}}{\text{Summe der möglichen Aufgabenpunkte}} \right) * 100$$

Quiz 1: Aufgabenformen und Aufgabentypen

Die ersten anderthalb Seminarsitzungen aller 4 Proseminare bezogen sich auf das Thema Aufgabentypen und Aufgabenformen. Die Teilnahme an diesen Sitzungen war verpflichtend und die Studierenden waren fast vollständig erschienen. Der Seminarleiter gestaltete diese Sitzungen überwiegend dozierend. Hierbei bemühte er sich, einen möglichst gleichartigen Unterricht in allen 4 Seminaren zu halten. Die Thematik wurde sehr gründlich und ausführlich dargelegt und durch eine umfangreiche Powerpoint-Folie unterstützt, die den Studierenden unmittelbar nach der Seminarsitzung zur Verfügung stand. In dieser Folie selbst waren auch 2 Aufgaben enthalten, wie sie im Quiz vorkommen könnten. Lerngrundlagen des Quiz Aufgabenformen waren demnach die Seminarsitzungen sowie die Powerpointfolie, jedoch keinerlei darüber hinausgehende Literatur.

Zum Themenbereich Aufgabentypen und Aufgabenformen wurden zwei Quiz konzipiert: das experimentelle Quiz, im folgenden Quiz 1 genannt und das Kriteriumsquiz oder Parallelretestquiz, welches die Auswirkungen des experimentellen Quiz erfassen sollte. Beide Quiz erfassten neben einigen Wissensaspekten vornehmlich Verständnis und Anwendung. Hierbei kamen verschiedene Aufgabentypen zum Einsatz (Short-Answer, MC mit einer oder mehreren korrekten Antworten, Zuordnungsaufgaben sowie einige Serien von True-false-Aufgaben.) Die Aufgabenkonstruktion des Parallelretest- bzw. Kriteriumsquiz orientierte sich an den Aufgabenstellungen des Quiz 1. Seine Items ähneln den analogen Items des Quiz 1 in unterschiedlichem Ausmaß. Neben identischen Items wurden im Retestquiz ähnliche, unterschiedliche Aufgaben, sowie auch einige Aufgaben mit einem etwas höherem Transfer erstellt, die insgesamt das Ziel verfolgten, zwei informelle zuverlässige Paralleltests zu erhalten, welche das angestrebte Lehrziel mit annähernd gleicher Validität erfassten. Quiz 1 Aufgabenformen diente als experimentelle Variante. Mit dem Parallelretestquiz Aufgabenformen sollte der Behaltenserfolg nach ca. einer Woche gemessen werden. Eine Überprüfung aller verfügbaren Daten ergab folgende Testkennwerte

Checkliste der Quiz Aufgabenformen:

Quiz 1: Aufgabenformen	: Aufgabenanzahl: 21	$\alpha = .78$	N=108
Quiz: Retest/Kriterium	: Aufgabenanzahl: 22	$\alpha = .76$	N=114
	Parallel-Retestreliabilität	$r_{tt} = .64$	N=103
	Messabstand 4-7 Tage		

Quiz 2: Skalen und Stichproben sowie Aufgaben stellen und Feedback geben

Im weiteren Seminarverlauf verfolgten die zwei Proseminare "Diagnostik Grundlagen" andere Lehrziele als die beiden verbliebenen Seminare "Diagnostik Aufgaben", was letztlich auch zu inhaltlich unterschiedlichen Quiz bei der nächsten Überprüfung führte. Die Studierenden der Seminare Diagnostik Grundlagen nahmen am Quiz Skalen und Stichproben teil, während die Studierenden der verbleibenden Seminare das Quiz "Aufgaben und Feedback" bearbeiteten. Der vorausgehende Unterricht gestaltete sich ähnlich wie beim Quiz Aufgabenformen als ein Dozieren des Seminarleiters auf der Basis einer den Studierenden verfügbaren Powerpointfolie. Eine Überprüfung aller verfügbaren Daten ergab folgende Kennwerte für die beiden Quiz

Quiz 2a: Aufgaben und Feedback	: Aufgabenanzahl: 29	$\alpha = .83$	N=53
Quiz 2b: Skalen und Stichproben	: Aufgabenanzahl: 15	$\alpha = .85$	N=58

Abschlussquiz für Quiz 1 und Quiz 2

Um die längerfristigen Auswirkung der Treatments zu testen, bearbeiteten die Studierenden in der letzten Seminarsitzung eine unbenotete Abschlussklausur, die etliche Aufgaben der vorangegangenen Quiz teils in identischer Form, teils als Parallelversionen umfassten. Hierbei orientierte sich die Aufgabenauswahl vornehmlich an hoher Trennschärfe und Differenzierungsfähigkeit zwischen den experimentellen Bedingungen, die auf der Basis der bearbeiteten Quiz 1 und 2 ermittelt werden konnten. Da in der Abschlussklausur aber weitere Themen anstanden, fielen die Anzahl der Aufgaben und damit die Zuverlässigkeit der Abschlussquiz(untertests) erwartungsgemäß geringer aus.

Abschlussquiz 1:	Aufgabenformen:	Aufgabenanzahl: 12	$\alpha = .65$	N=106
Abschlussquiz 2a:	Aufgaben und Feedback:	Aufgabenanzahl: 10	$\alpha = .53$	N=49
Abschlussquiz 2b:	Skalen und Stichproben:	Aufgabenanzahl: 7	$\alpha = .57$	N=58

Versuchsplan

Die Studierenden wurden zu Beginn des Seminars nach Zufall auf die Gruppen A und B zugeteilt. Hierzu führte der Versuchsleiter die Zufallsprozedur für jedes Seminar gesondert durch, um sicher zu stellen, dass sich die Studierenden in jedem Seminar zur Hälfte auf die beiden Gruppen verteilen. Durch diese stratifizierte Zufallszuweisung wird die Vergleichbarkeit der Studierenden in beiden Gruppen ziemlich optimal gewährleistet, da mögliche Unterschiede in den Seminaren in jeder Bedingung angemessen repräsentiert sind. Tabelle 2 schematisiert den Versuchsplan. Für jede Gruppe liegen 4 Messungen vor. Das fett gedruckte benotet und unbenotet charakterisiert eine Messung unter einer

bestimmten experimentellen Quizbedingung, die stets angekündigt war, während der Parallelretest und das Abschlussquiz überraschend vorgelegt wurden und den Studenten dort vor der aktuellen Bearbeitung bewusst war, dass sie nicht benotet werden würden.

Tabelle 2: Verdeutlichung des Versuchsplans

	10.11.2008	17.11.2008	1.12.2008	9.2.2009
	Quiz 1	Parallelretest zu Quiz 1	Quiz 2 Abschlussquiz zu Quiz 1 und Quiz 2
R Gruppe A:	benotet	unbenotet	unbenotet	unbenotet
R Gruppe B:	unbenotet	unbenotet	benotet	unbenotet

Der erste Vergleich der Gruppen A und B beim Quiz 1 entspricht einem klassischen Randomisierungsplan. Im Quiz 1 wird die unmittelbare Wirkung einer Benotungsandrohung auf die aktuelle Quizleistung, im unangekündigten Parallelretest die etwas langfristige Auswirkung der Bedingungen unter Quiz 1 erfasst. Danach vollzogen beide Gruppen einen Bedingungswechsel, dessen unmittelbare Wirkung in Quiz 2 ermittelt wurde. Die Fragestellung lässt sich somit zunächst an einem für alle Beteiligten identischen Lehrstoff auf höchst möglichem Niveau interner Validität prüfen. Der Treatmentreversal ermöglicht eine erneute Testung der Hypothese an einem anderen Lehrstoff. Im Abschlussquiz wird schließlich eine längerfristige Behaltensprüfung der beiden Quizbedingungen für beide Quiz angestrengt.

Die Studierenden der Gruppe A und B unterscheiden sich nicht signifikant hinsichtlich Alter, Geschlecht und Abiturnotendurchschnitt. Der besonders wichtige Abiturnotendurchschnitt fällt für beide Gruppen mit 2,40 und 2,37 nahezu identisch aus. Alle Ergebnisse stützen die Effizienz der Randomisierung.

Durchführung der Quiz 1 und 2

Die Studierenden bearbeiteten das benotete Quiz 1 in der Seminarsitzung und erhielten nach Einsammlung der Quizhefte Rückmeldung über die korrekten Antworten. Das unbenotete Quiz fand um ein bis zwei Tage verzögert zu Hause am Computer statt. Es musste innerhalb von 2 Tagen in Angriff genommen werden. In der nächsten verpflichtenden Seminarsitzung teilte der Seminarleiter für alle gänzlich unerwartet die Parallelversion des Quiz 1 aus und ließ es analog der Prozedur beim benoteten Quiz 1 bearbeiten. Die Quizhefte dieses Retestquiz wurden anschließend eingesammelt und auf jegliches Feedback verzichtet. Die Durchführung der nachfolgenden Quiz 2a und b entsprach dem Vorgehen bei Quiz 1 mit dem Unterschied, dass die Bedingungen für die Gruppen nun vertauscht wurden.

Ergebnisse

Auswirkung der Quizbedingung auf die Leistung en im Quiz

Alle Studierenden bearbeiteten sowohl ein benotetes wie ein unbenotetes Quiz. Zunächst werden nur die Ergebnisse dieser Quiz dargestellt. Daraus wird erkennbar, welche Wirkung die Erwartung einer Benotung bzw. Nichtbenotung auf das Ergebnis im entsprechenden Quiz ausübt.

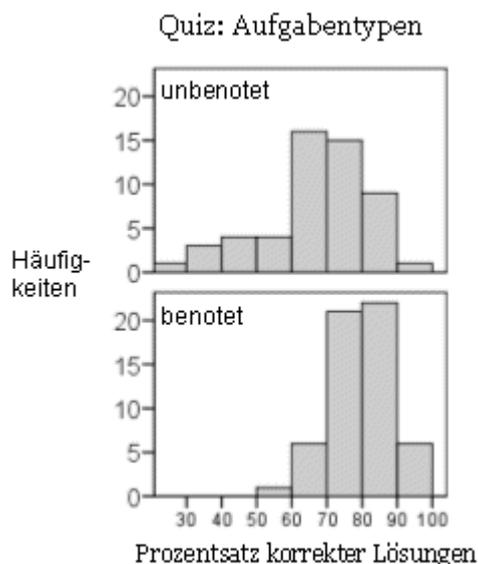
Quiz 1: Aufgabentypen und Aufgabenformen

Der Bedingungsvergleich beim ersten Quiz entspricht einem klassischen Randomisierungsplan ohne Vortest. Alle Studierenden aller Seminare bearbeiteten hierbei dasselbe Quiz. Tabelle 3 stellt die deskriptiven Ergebnisse sowie das Ergebnis des Mittelwertvergleichs beider Bedingungen dar.

Tabelle 3: Prozentsatz korrekter Lösungen in Quiz 1

	M	s	N	t	p	d
Quiz benotet:	80,1	7,5	56			
Quiz unbenotet:	70,1	14,5	52	4,4	0.001	.87

Wie man der Tabelle 3 entnehmen kann, erzielten die Studierenden unter der Bedingung benotetes Quiz mit durchschnittlich 80% korrekter Lösungen 10% bessere Ergebnisse als die Studierenden unter der Bedingung unbenotetes Quiz. Der mittels t-Tests für unabhängige Stichproben durchgeführte Mittelwertsunterschied überbietet das geforderte Signifikanzniveau von 1 Prozent und entspricht einer Effektstärke von $d = .87$ (95%iges Vertrauensintervall für d : .47-1.26) zugunsten des benoteten Quiz. Auch die Varianzen beider Gruppen unterscheiden sich hochsignifikant ($F=15,9$), was auch in den Histogrammen der Abbildung 1 zum Ausdruck kommt.

Abbildung 1: Histogramme der Leistungsergebnisse unter den beiden Quizbedingungen des Quiz 1.

Während die Daten unter der Bedingung eines benoteten Quiz symmetrisch in Form einer Normalverteilung relativ eng und dem Mittelwert schwanken, fällt beim unbenoteten Quiz der linke Teil der Verteilung besonders auf. Die schwachen Ergebnisse deuten darauf hin, dass etliche Studierende das Quiz nicht so ernst genommen haben bzw. vermutlich schlecht vorbereitet zum Quiz angetreten waren. Weiterhin findet man unter der Bedingung unbenotetes Quiz weniger Leistungsspitzen, was bei einigen Studenten Zweifel an der Bereitschaft aufkommen lässt, sie hätten ihr mögliches Leistungspotenzials voll ausgeschöpft.

Quiz 2 mit Bedingungswechsel

Im nächsten Quiz wurden die Quizbedingungen für die Studierenden vertauscht. Wer zuvor das benotete Quiz absolvierte, musste nun ein unbenotetes Quiz bearbeiten und umgekehrt. Da in den Seminaren zwischenzeitlich unterschiedliche Themen behandelt wurden, sind die Ergebnisse in Tabelle 4 nach den jeweils durchgeführten Quizthemen getrennt dargestellt und statistisch verarbeitet worden.

Tabelle 4: Prozentsatz korrekter Lösungen im Quiz 2:

	M	s	N	t	p	d
a) Quiz 2a: Aufgaben stellen und Feedback geben						
benotet:	75,5	12,2	28			
unbenotet:	63,6	14,3	25	3.3	0.01	0.90
b) Quiz 2b : Skalen und Stichproben						
benotet:	70,0	15,1	24			
unbenotet:	54,9	23,4	29	3.0	0.01	0.77

Die Ergebnisse entsprechen konsistent den früheren Befunden. Studierende, die ein benotetes Quiz bearbeiteten, erzielten in diesem Quiz signifikant bessere Ergebnisse als Studierende, die dasselbe Quiz ohne Benotungserwartung durchführten. Der Leistungsvorteil für das benotete Quiz liegt stets bei mindestens 10%, was einer Effektstärke hoher praktischer Bedeutsamkeit entspricht.

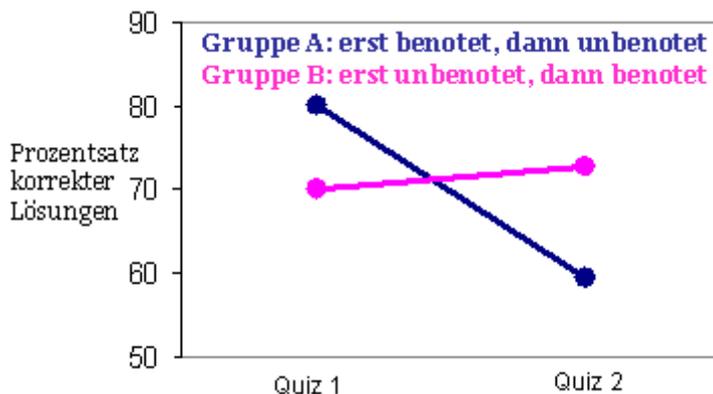
Zusammenfassende Darstellungen der Leistungsdaten unter den Quizbedingungen.

Auch wenn die bisherigen Ergebnisse bereits hinreichende Evidenz zur Leistungsüberlegenheit angekündigt benoteter Quiz hinterlassen, soll im folgenden dieser Vorteil noch etwas übersichtlicher herausgestellt werden. Abbildung 2a veranschaulicht die Mittelwerte der Gruppen A und B in beiden Quiz. Die Mittelwerte beim Quiz 2 ergeben sich hierbei aus den Mittelwerten von Quiz 2a und 2b. Man erkennt, dass Quiz 2 insgesamt schwieriger war. In Abbildung 2b wurden die Ergebnisse für jedes Quiz z-transformiert. Da im Quiz 2 unterschiedliche Themen getestet wurden, dient als z-Mittelwert des Quiz 2 jeweils der nach Probandenanzahl gewichtete z-Mittelwert aus Quiz 2a und 2b

Abbildung 2

Zusammenfassende Darstellung der Ergebnisse in benotetem und unbenoteten Quiz

a) Mittelwerte der Gruppen in beiden Quiz



b) z-Mittelwerte der Gruppen in beiden Quiz

benotet; unbenotet

	Quiz 1	Quiz 2
Gruppe A	.39	-.39
Gruppe B	-.42	.37

In Abbildung 2a und b zeigt sich der Leistungsvorteil der benoteten Quiz konsistent sowohl querschnittlich im direkten Vergleich der Gruppen untereinander, wie auch längsschnittlich betrachtet in der Leistungsentwicklung einer Gruppe nach einem Bedingungswechsel. Die Unterschiede fallen deutlich und in vergleichbarer Höhe aus. Alle möglichen Vergleiche unterstützen eindeutig die These, die Erwartung einer Quizbenotung bewirke eine höhere Leistung im entsprechenden Quiz als die Erwartung, das Quiz würde nicht benotet.

Behaltensleistung bei Quiz 1: Aufgabentypen und Aufgabenformen: unerwartet in der nächsten Seminarsitzung

Der Parallelretest zu Quiz 1 wurde für alle Studierenden unerwartet in der dem Quiz 1 folgenden Seminarsitzung zur Bearbeitung vorgelegt, um die Stabilität des Wissens zu prüfen. Der Zeitabstand betrug für die Studierenden unter benoteter Quizbedingung exakt 7 Tage, während er bei den Studierenden des unbenoteten Onlinequiz in Schnitt ca. 2 Tage kürzer ausfiel und insgesamt zwischen 4 und 7 Tagen schwankte. Im Gegensatz zum ersten Quiz Aufgabenformen bearbeiteten die Studierenden den Retest Aufgabenformen exakt unter den gleichen kontrollierten Bedingungen im Seminarraum. Vor der Testung war angekündigt worden, das Ergebnis des Retest gehe nicht in die Seminarnote ein, stelle aber eine sinnvolle Übung zur Stabilisierung des Wissens dar.

Tabelle 5: Prozentsatz korrekter Lösungen im unbenoteten Parallelretest Aufgabenformen.

Bedingung	M	s	N	t	p	d
Quiz 1 benotet:	78.5	9.0	54			
Quiz 1 unbenotet:	66.3	14.0	49	5.3	0.001	1.04

Die Ergebnisse entsprechen weitgehend den Befunden zu Quiz 1. Studierende, die ihr erstes Quiz unter Benotungserwartung absolvierten, erzielten im Parallelretest signifikant und mit einer Effektstärke von $d=1.04$ [95%iges Vertrauensintervall von $d: .62-1.45$] praktisch bedeutsam bessere Testergebnisse als die Studierenden unter vormals unbenoteter Quizbedingung. Die Bearbeitung des ersten Quiz unter der vorherigen Ankündigung, dieses Quiz zu benoten, bewirkte folglich eine deutlich höhere Lernaneignung als die Quizbearbeitung ohne erwartete Notenkonsequenz. Diese höhere Lernaneignung blieb wenigstens ca. eine Woche lang mindestens vollständig aufrecht erhalten. Der deutliche Leistungsvorteil im Retest für die Studierenden unter vorheriger Quizbenotung stützt die These, ein benotetes Quiz mit KCR-Feedback nach dem Quiz verbessere etwas langfristigeres Lernen deutlich besser als ein verpflichtendes unbenotetes Quiz mit vielfältigen Rückmeldemöglichkeiten.

Behalten im Abschlussquiz unerwartet in der letzten Seminarsitzung

Ohne vorherige Ankündigung fand in der letzten Seminarsitzung die unbenotete Abschlussklausur statt. Um die Testmotivation zur Bearbeitung der unbenoteten Abschlussklausur möglichst hoch zu halten, wurden unter anderem leistungsabhängige Geldprämien eingesetzt (siehe Jacobs 2009). Die Abschlussklausur selbst beinhaltete Aufgaben der vorangegangenen Quiz in identischer sowie leicht modifizierter, paralleler Form. Der Zeitabstand vom jeweiligen Quiz bis zur Abschlussklausur beträgt ca. 6 bis 8 Wochen. Im Folgenden werden die Ergebnisse nach den unterschiedlichen Seminaren getrennt mitgeteilt. Gruppe A bearbeitete Quiz 1 als benotete und Quiz 2 als unbenotete Version, während Gruppe B zunächst das unbenotete und dann das benotete Quiz 2. absolvierte. Sollte die Benotung eines Quiz die Leistung nachhaltig verbessern, so wären entsprechende höhere Testwerte bei den Aufgaben der Abschlussklausur zu erwarten, die denen der benoteten Quiz entsprechen.

Tabelle 6 zeigt die Gruppenmittelwerte für jedes Quiz und das jeweilige Ergebnis der Gruppenmittelsunterschiede im T-test für unabhängige Stichproben. Alle Ergebnisse liegen numerisch, zwei signifikant in erwarteter Richtung..

Der Gesamteffekt der Quizbenotung zeigt sich sehr deutlich und umfassend in der hochsignifikanten Interaktion zwischen Gruppenzugehörigkeit (A, B) und Quizfolge (Quiz 1 und 2). Diese Berechnung basiert auf der zweifaktoriellen Varianzanalyse mit dem Between-Faktor Gruppe und dem Messwiederholungsfaktor Quiz. Die Interaktion erfasst die unterschiedlichen Quizbedingungen der Gruppen im Zeitverlauf der beiden Testungen. Das Resümee dieser Analyse führt zu der eindeutigen Schlussfolgerung, benotete Quiz bewirkten höhere Leistungen, die mindestens 6 bis 8 Wochen überdauern und vom Ausmaß her insgesamt einer mittleren Effektstärke entsprechen.

Tabelle 6: Ergebnisse der Abschlussklausur für identische und parallele Aufgaben aus benotetem und unbenotetem Quiz

a) Seminare Diagnostik Aufgaben: **benotet**; unbenotet

		Abschlussklausur			
		Aufgaben zu			
		Quiz 1		Quiz 2a	
		[$\alpha=.64$]		[$\alpha=.53$]	
		M	<i>s</i>	M	<i>s</i>
Gruppe A :		72.5	9.2	58.7	16.2
Gruppe B :		63.6	15.1	64.7	13.4
		t(47)= 2.43		t(47)= -1.41	
		p _e = 0.01		p _e = 0.08	
		d = .71		d = .40	

Interaktion zwischen Gruppe und Quiz: $F(1,47)=12.81$ $p=.001$

b) Seminare Diagnostik Grundlagen **benotet**; unbenotet

		Abschlussklausur			
		Aufgaben zu			
		Quiz 1		Quiz 2b	
		[$\alpha=.67$]		[$\alpha=.77$]	
		M	<i>s</i>	M	<i>s</i>
Gruppe A :		68.1	11.4	54.8	22.0
Gruppe B :		63.5	13.9	66.2	19.6
		t(55)= 1.35		t(55)=-2.19	
		p _e = 0.09		p _e = 0.016	
		d = .36		d = .55	

Interaktion zwischen Gruppe und Quiz: $F(1,55)= 14.4$ $p<.001$

Subjektive Daten

Aktuelle Angst

Unmittelbar vor einem Quiz füllten die Studenten den aktuellen Angstfragebogen Situative aktuelle Prüfungsangst (SPA) sowie das Fearthermometer (FT) aus (siehe Jacobs 1996). Der SPA lag die Anweisung zugrunde: "Sie erhalten gleich das Quiz zur Bearbeitung." Durch die Angsterhebung sollte geprüft werden, ob benotete Quiz höhere Angst auslösen als unbenotete, aber bewertete Quiz. Wenn mehr auf dem Spiel steht, ist in der Regel auch eine höhere Angst zu erwarten. Nach Jacobs (1981) entsteht aktuelle Angst durch die Kombination aus Misserfolgsbefürchtung und dem Ausmaß der mit dem Misserfolg verknüpften aversiven Konsequenzen, die sich recht gut aus der subjektiven Bedeutsamkeit des Leistungsergebnisses ableiten lassen. Da die Misserfolgserwartung in einer Prüfung häufig nicht hinreichend ausgeschlossen werden kann, gewinnt die subjektive Bedeutsamkeit des angestrebten Leistungsergebnisses bzw. die erwartete Aversivität einer möglichen Zielverfehlung besondere Relevanz für die Angstauslösung einer Prüfung. Da das Ausmaß der Aversivität einer persönlichen Zielverfehlung in einer benoteten Prüfung höher ausfällt, ist im Mittel eine höhere Angst in einer benoteten Prüfung zu erwarten. Wenngleich die theoretischen Aussagen diese Hypothese eindeutig vermuten lassen, sind die bisherigen empirischen Ergebnisse von Jacobs keineswegs eindeutig. Sie sprechen in der Tendenz zwar für eine etwas höhere aktuelle Angst unter einer benoteten Testbedingung, die aber statistisch schwer zu belegen ist. Außerdem erschwert die schwache Versuchplanung der bisherigen Studien die Aussagekraft der Befunde. So fand Jacobs (2003a) keine signifikanten Angstunterschiede unmittelbar vor einer anonymen Probeklausur und der eine Woche später folgenden echten benoteten Klausur. Vorteil dieser Studie ist es, den Unterschied zwischen benoteten und unbenoteten Tests auf versuchsplantechnisch höchstem Niveau durchführen zu können, so dass die Hypothese zumindest für die hier vorliegenden Quizbedingungen eindeutig entschieden werden kann.

Wie aus Tabelle 7 hervorgeht, fielen die aktuelle Angst und das Fearthermometer unmittelbar vor dem benoteten Quiz stets signifikant höher aus als die Angst vor dem unbenoteten Quiz. Das Ausmaß der Unterschiede liegt im mittleren, teilweise sogar im hohen Effektstärkebereich.

Tabelle 7: Ergebnisse zur aktuellen Angst in benotetem und unbenotetem Quiz

	Quiz	Reli	benotet		unbenotet		t	p	d
			M	s	M	s			
aktuelle Prüfungsangst SPA	1	.92	25.7	8.1	20.5	10.7	2.8	<.01	0.55
	2a	.95	25.6	8.9	19.8	11.3	1.9	<.05	0.58
	2b	.94	25.6	10.7	17.2	8.2	3.2	<.01	0.88
Fearthermometer FT	1	.83	3.1	2.1	2.2	1.8	2.4	<.01	0.47
	2a	.77	3.1	1.9	1.7	1.9	2.4	<.05	0.76
	2b	.85	2.8	2.3	1.6	1.9	2.1	<.05	0.57

	N	N
	benotet	unbenotet
Quiz 1: Aufgabenformen und Aufgabentypen	56	40-51
Quiz 2a: Aufgaben stellen und Feedback geben	24	25
Quiz 2b: Skalen und Stichproben	28-29	25-29

Als Reliabilität wurde bei der SPA Conbachs α berechnet. Beim FT dient die Korrelation zwischen FT und SPA als untere Grenze der Reliabilität. Das den Reliabilitäten zugrunde liegende N ergibt sich aus der Summe von N benotet und N unbenotet.

Gemessen an den bisherigen Ergebnissen von Jacobs (1996..2005), fallen die ungewöhnlich konsistenten SPA-Angstmittelwerte von ca. 26 im benoteten Quiz aber noch erkennbar - grob betrachtet ca. 0.5 Standardabweichungen - geringer aus als bei Klausuren, was höchstwahrscheinlich darauf zurück geht, dass die Gewichtung der Quiz für die Seminarendnote schwächer ausfällt als bei einer Abschlussklausur.

Abbildung 3: Angstunterschiede der Quizbedingungen und Angstentwicklungen beim Bedingungswechsel

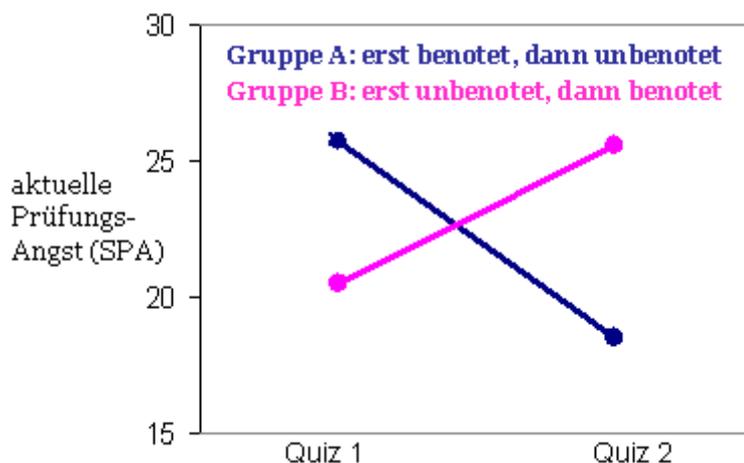


Abbildung 3 stellt die Ergebnisse zur aktuellen Prüfungsangst zusammenfassend dar und zeigt anschaulich auf, dass die Hypothese höherer Angst in benoteten Quiz bei allen möglichen Vergleichen und damit letztendlich idealtypisch bestätigt werden konnte.

Vorbereitungsintensität, Anstrengung und Bedeutsamkeit des Quiz

Nach Quizabgabe sollten die Studierenden einige Fragen zur Quizvorbereitung, Anstrengungsbereitschaft und der subjektiven Bedeutsamkeit des Quiz beantworten. Beim ersten Quiz fehlten einige Fragen so dass die Ergebnisse für manche Variablen nur beim Quiz 2 in Erscheinung treten konnten. Tabelle 8 legt dar, mit welchen Fragen die einzelnen Variablen erhoben wurden. Jeder Frage war eine 5 stufige, auf jeder Stufe verbal verankerte Likert-Skala unterlegt. Die Beantwortungsmöglichkeit

reichte von "stimmt überhaupt nicht" (=1) bis "stimmt ganz genau" (=5). Als Variablenwert dient der Durchschnitt aus den in die richtige Richtung gepolten Items.

Tabelle 8: Eingeschätzte Quizvorbereitung, Anstrengungsbereitschaft und Bedeutsamkeit

Name	Items
subjektiv eingeschätzte Quizvorbereitung	1.) Ich hatte mich vor dem Quiz wenig mit dem Thema Skalen und Stichproben befasst. (-) 4.) Ich hatte die Folien des Seminars zu Skalen und Stichproben durchgearbeitet. 6.) Ich hatte mich auf die Themen und Inhalte des Quiz gut vorbereitet.
Zeit für Quizvorbereitung	16.) Ich schätze die für das Quiz investierte Vorbereitungszeit auf ? Minuten.
investierte Anstrengungsbereitschaft	2.) Ich habe alle meine Kräfte eingesetzt, um das Quiz aufmerksam zu bearbeiten. 5.) Ich habe das Quiz voll konzentriert mit höchster Anstrengung bearbeitet. 7.) Ich sah keine Veranlassung, mich im Quiz ganz besonders anzustrengen. (-)
Bedeutsamkeit guter Leistungen	3.) Es ist für mich sehr wichtig, in diesem Quiz gut abzuschneiden. 8.) Das Ergebnis im Quiz hat für mich keine besondere Bedeutung. (-)

Die theoretischen Erwartungen lagen offen auf der Hand. Die Studierenden sollten besser vorbereitet zum benoteten Quiz antreten, sich während der Quizbearbeitung möglichst stark anstrengen und besonderen Wert auf ein gutes Ergebnis legen. Alle möglichen Befunde in Tabelle 9 stützen diese Hypothesen und erbrachten Unterschiede im hohen bis sehr hohen Effektstärkebereich.

Tabelle 9: Ergebnisse zu den subjektiven Erhebungen.

	Quiz Reli	benotet		unbenotet		t	p	d	
		M	s	M	s				
Quizvorbereitung	1	.88	4.1	.9	3.1	1.0	5.3	<.01	1.10
	2a	.74	4.1	.7	3.0	.9	4.7	<.01	1,32
	2b	.81	4.3	.6	2.9	.9	7.6	<.01	2.00
Zeit für Quizvorbereitung in Minuten	2a		135.8	69.0	35.8	27.4	6.7	<.01	1.90
	2b		74.8	49.7	24.2	22.3	5.0	<.01	1.31
Anstrengungsbereitschaft	1	.89	4.6	.5	3.8	.8	5.6	<.01	1.13
	2a	.60	4.4	.7	3.7	.6	3.7	<.01	1.05
	2b	.66	4.5	.5	3.8	.7	4.7	<.01	1.24
Bedeutsamkeit	2a	.73	4.0	.8	3.1	1.0	3.5	<.01	1.00
	2b	.73	4.4	.5	3.0	.6	8.9	<.01	2.18

	N	N
	benotet	unbenotet
Quiz 1: Aufgabenformen und Aufgabentypen	56	40-51
Quiz 2a: Aufgaben stellen und Feedback geben	24	25
Quiz 2b: Skalen und Stichproben	28-29	25-29

Als Reliabilität wurde Conbachs α berechnet. Das den Reliabilitäten zugrunde liegende N ergibt sich aus der Summe von N benotet und N unbenotet.

Als wichtigste Einflussgröße für den Lernerfolg betrachte ich die investierte Zeit zur Erarbeitung des Lehrstoffs, vorausgesetzt, der Studierende nutze die quizrelevanten Informationsquellen zielgerichtet und lernstrategisch angemessen. Der Lerneinsatz kann natürlich nur dann die Lernleistung fördern, wenn die Quizanforderungen die geistigen Kapazitäten und verfügbaren Ressourcen nicht überfordern. Gegen eine Überbeanspruchung sprechen einerseits die insgesamt recht guten Leistungsergebnisse, wie auch die subjektiv empfundene Anstrengung der Studierenden. So lag der Mittelwert bei allen Quiz unter allen Bedingungen für die Behauptung "Das Quiz war ziemlich anstrengend" auf der 5 stufigen Likertskala im mittleren Bereich bei 3 (=weder/noch). Auch wenn im Quiz wenig Faktenwis-

sen, sondern meistens Verstehens- und Anwendungsaufgaben gestellt wurden, so waren diese Lehrziele durch intensives Studium gut zu erarbeiten. Sowohl die subjektiv eingeschätzte Quizvorbereitung wie die eingeschätzte Vorbereitungszeit fallen beim benoteten Quiz mit Effektstärken von jeweils über 1 sehr überzeugend höher aus. Die eingeschätzte Vorbereitungszeit auf ein benotetes Quiz liegt mindestens 3 mal so hoch wie bei einem unbenoteten Quiz. Diese deutlichen Unterschiede in der Vorbereitungszeit auf die Quiz entsprechen den Befunden von Narloch, Garbin & Turnage (2006)

Auch die im Quiz investierte Anstrengungsbereitschaft sowie die Bedeutsamkeit, ein gutes Ergebnis zu erzielen, fielen bei allen benoteten Quiz deutlich höher aus. Da das Quizergebnis meiner Meinung nach im Wesentlichen von der Vorbereitung abhängt, kann eine verstärkte Bemühung während der Quizbearbeitung weit weniger Lernzuwachs erbringen. Recht schwaches Leistungsverhalten der Kontrollgruppe könnte aber durchaus einen Leistungsunterschied bewirken, der in etlichen Studien auch gefunden wurde. (siehe zusammenfassend Jacobs 2007, Kap: Testmotivation und Testleistung bei benoteten und unbenoteten Examen). Hier erzielten aber auch die Studierenden unter unbenoteter Quizbedingung relativ passable Mittelwerte ihrer investierten Anstrengungsbereitschaft. Es klingt theoretisch plausibel anzunehmen, die erhöhte Bedeutsamkeit des Leistungsergebnisses unter erwarteter Benotung sei die entscheidende Ursache sowohl für die bessere Vorbereitung auf das Quiz wie auch für die deutlich höhere investierte Anstrengungsbereitschaft bei der Quizbearbeitung.

Die Wirkung unbenoteter Quiz auf die Ergebnisse der Abschlussklausur

Gruppe A und B der Seminare Diagnostik Grundlagen absolvierten weitere Quiz, unter anderem jeweils ein unbenotetes Onlinequiz (abgekürzt U). Gruppe A bearbeitete recht früh das Quiz U1 "Ziele und Strategien pädagogischer Diagnostik" und Gruppe B ziemlich am Ende des Seminars das Quiz U2 "Aufgabenanalyse". Diese Quiz kamen ausschließlich als verpflichtende, unbenotete Onlinequiz zum Einsatz. Dadurch sollte geprüft werden, ob die verpflichtende, aber nicht mit Noten bewertete Bearbeitung von Onlinequiz mit vielfältigen Rückmeldungen den Lernerfolg langfristig fördert. Es steht folglich die Frage im Raum, ob die Bearbeitung eines Onlinequiz mit hinreichenden Rückmeldungen überhaupt einen Lern- bzw. Behaltenseffekt nach sich zieht und letztlich besser als gar nichts einzuschätzen ist. Während der Dozent den Lehrstoff zu Quiz U1 analog der bisherigen Vorgehensweise dozierend in einer Seminarsitzung präsentierte, fehlte die Zeit, um Analoges auch für das Thema Aufgabenanalyse zu realisieren. Stattdessen machte er allen Studierenden eine Powerpoint-Folie zum Thema Aufgabenanalyse zugänglich, die es zu Hause zu bearbeiten galt. Gruppe B musste dann das Onlinequiz Aufgabenanalyse bearbeiten, Gruppe A hatte keinen Zugang zu diesem Quiz. Tabelle 10 erlaubt einen Mittelwertsvergleich der entsprechenden Aufgaben in der Abschlussklausur von Studierenden, die am entsprechenden verpflichtenden unbenoteten Quiz teilnahmen mit denen, die kein derartiges Quiz bearbeiteten.

Tabelle 10: Ergebnisse der Abschlussklausur für identische und parallele Aufgaben aus Quiz U1 und U2

unbenotetes Quiz; **kein Quiz** (U1=Ziele und Strategien, U2=Aufgabenanalyse)

		Lehrziele der Abschlussklausur aus			
		Quiz U1		Quiz U2	
		[$\alpha=.24$]		[$\alpha=.48$]	
		M	<i>s</i>	M	<i>s</i>
Gruppe A :		65.2	16.5	40.1	26.5
Gruppe B :		64.2	16.3	54.7	22.3
		t(56)= 0.10		t(56)=-2.27	
		p _e = 0.46		p _e = 0.014	
		d = .02		d = .60	

Interaktion zwischen Gruppe und Quiz: $F(1,56) = 6,202$, $p = 016$

Die den Gesamteffekt am besten erfassende Interaktion zwischen Gruppe und Quiz bestätigt den behaltenswirksamen Einsatz des verpflichtenden Onlinequiz. Überzeugend zeigt sich dieser Effekt allerdings nur beim Quiz Aufgabenanalyse, was an der Lehrvermittlung (nur Powerpoint), am Inhalt

bzw. der Schwierigkeit der Thematik (z.B. einige schwierigere Anwendungsaufgaben) liegen oder auch durch das relativ kurze Zeitintervall von lediglich ca. 2.5 Wochen zur Abschlussklausur begünstigt sein könnte. Für die Wirkungslosigkeit des Onlinequiz beim Quiz Diagnostische Strategien finde ich hingegen keine Erklärung. Wegen der sehr geringen Aufgabenanzahl der Quiz von jeweils 5 in der Abschlussklausur war keine voll befriedigende Prüfung der Hypothese möglich. Wenn sich dann doch zumindest ein positiver Effekt finden lässt, so spricht dieser für die These, auch unbenotete Quiz mit KCR bzw. elaborem Feedback könnten durchaus einige Lern- und Behaltenseffekte zeigen.

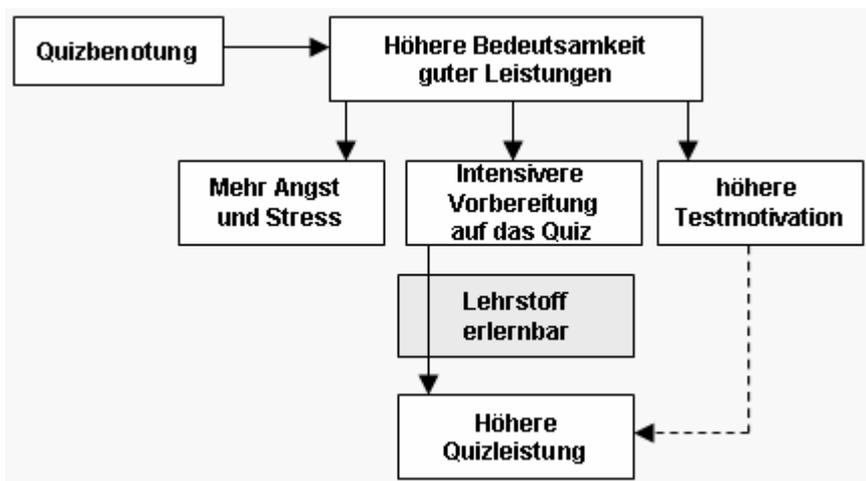
Zusammenfassung und Diskussion

Die wichtigsten Ergebnisse und einige Interpretationshinweise

Die Studie bestätigte in vollem Umfang die These, benotete Quiz erzielten höhere Testleistungen als verpflichtende unbenotete Quiz mit ausführlichem Feedback. Wegen der Trivialität der Fragestellung mag das Ergebnis nicht besonders überraschen. Schon weniger trivial schätze ich ein, dass der durch Notenandrohung bewirkte deutliche Lerneffekt mindestens eine Woche voll erhalten blieb und sich im Ausmaß mittlerer Effektstärke auch noch 6 bis 8 Wochen später eindeutig belegen ließ. Im Vergleich mit einer hypothetischen No-Treatment- Kontrollgruppe stellt das hier gefundene Ausmaß des Leistungsvorteils benoteter Quiz eher eine Unterschätzung dar. Zudem konnte die Lernwirksamkeit von unbenoteten Onlinequiz mit ausführlichen Rückmeldungen gegenüber gar nichts, wenn auch vom Ausmaß her relativ schwach, so doch auf versuchsplantechnisch hohem Niveau zumindest teilweise statistisch belegt werden.

Die Arbeit liefert weiterhin etliche Stützen für die Erklärung, die Überlegenheit des benoteten Quiz mit der höheren Lernmotivierung und dem daraus resultierenden höheren Lerneinsatz zu begründen. Aus motivationaler Sicht wurde die Erwartungs-mal-Wert-Theorie empirisch eindrucksvoll durch die klaren Unterschiede in der subjektiven Bedeutsamkeit des Leistungsergebnisses gestützt. Sie wird als wesentliche Ursache für die erhöhte Quizvorbereitung und Testmotivation bei benoteten Quiz angesehen. Die Erwartung, sein Ziel durch vermehrte Anstrengung auch erreichen zu können, wurde zwar nicht direkt erhoben, lässt sich aber stark vermuten und ist zudem realistisch, weil die geforderten Quizleistungen nicht hoch fähigkeitsabhängig, sondern überwiegend erlernbar waren. Deshalb führte die Erwartung der bedeutsameren Ergebniskonsequenz auch tatsächlich zu höherer Leistung. Die höhere Leistung benoteter Quiz wurde auch durch mehr Angst und Stress erkauft, was die Frage aufwirft, ob die Leistungssteigerung diese Belastung wirklich aufwiegt. Abbildung 4 fasst das vermutete Wirkungsgefüge der Variablen zusammen:

Abbildung 5: Durch Quizbenotung angeregte Erhöhung der Bedeutsamkeit guter Leistungsergebnisse und ihre Auswirkungen.



Eine Replikation an zwei weiteren Quiz der Seminare Diagnostik Grundlagen führte zu einer überzeugenden Bestätigung der hier gezeigten Befunde für die Testleistungen und die motivationalen Vari-

ablen. Unter anderem konnte dort der Leistungsvorteil unter Notenerwartung auch dann nachgewiesen werden, als alle Studierende das Quiz im Seminar unter kontrollierten Bedingungen absolvierten.

Höhere Lerneffektivität des benoteten Quiz durch vermehrte Vorbereitungszeit

Da die höhere Leistung im benoteten Quiz durch vermehrte Lernzeit erzielt wurde, sind benotete Quiz zwar lerneffektiver, aber nicht automatisch lerneffizienter als unbenotete Quiz. Im Hinblick auf die Frage, wie man besser lernt, muss der Methodenvergleich daher als unfair bezeichnet werden. Selbst wenn man ein unbenotetes Quiz sehr konzentriert und konstruktiv bearbeitet und die Rückmeldungen intensiv studiert, so reicht dieser potenzielle Lerngewinn aber offenbar nicht aus, um mit dem zuvor aktivierten höheren Lernengagement infolge der Bedrohungsinduzierung einer Benotung konkurrieren zu können. Weil die benötigte Lernzeit unter beiden Bedingungen durch die Randomisierung konstant gehalten war, bestätigen die Befunde im übrigen recht gut das recht abstrakte Modell schulischen Lernens von Carroll (1963): $\text{Lernerfolg} = (\text{aufgewendete Lernzeit} / \text{benötigte Lernzeit})$. Das mögliche Ausmaß des Benotungseffektes hängt vermutlich stark vom Anspruchsniveau bzw. der Schwierigkeit der Lehrziele und letztendlich von der Enge des Zusammenhangs zwischen Lerneinsatz und Lernergebnis ab.

In den hier untersuchten Seminaren war keine benotete Abschlussklausur vorgesehen und kein Studierender hat mit einer solchen gerechnet. Insofern erbringt hohes Wissen im unbenoteten Quiz keinerlei potenzielle "notenrelevante" Vorteile für die Zukunft, was das Engagement zur Vorbereitung auf ein unbenotetes Quiz vermutlich in Grenzen hielt. Möglicherweise wächst die Effektivität unbenoteter Quiz mit der Erkenntnis, frühzeitige Vorbereitung und gute Leistungen im unbenoteten Quiz verbesserten deutlich die Chancen in einer benoteten Abschlussklausur.

Warum unangekündigte unbenotete Kriteriumstests ?

Als kurz- und langfristige Behaltenskriterien wurden bewusst unangekündigte, unbenotete Tests eingesetzt. Deshalb lässt sich aus den Ergebnissen auch keineswegs folgern, in einer, der pädagogischen Praxis eher folgenden, benoteten Abschlussklausur wären ohne weiteres vergleichbare Vorteile für die benoteten Quiz herausgekommen. (siehe z.B. Azorlosa & Renner 2006). Ich vermute eher nicht oder zumindest geringere Vorteile, weil die Erwartung einer Benotung die meisten Studierenden dazu veranlasst hätte, sich entsprechend ihrem aktuellen Wissensstand möglichst gut auf diese Abschlussklausur vorzubereiten. Da die Studierenden in den unbenoteten Quiz mehr Lerndefizite aufzuweisen hatten, hätten Sie höchstwahrscheinlich die entsprechenden Wissensgebiete intensiver vorbereitet bzw. die zuvor unzureichende Vorbereitung dann nachgeholt. Narloch, Garbin, & Turnage (2006) fanden in ihrer Studie Befunde, die in diese Richtung weisen.

Umso erstaunlicher sind die hier gefundenen Ergebnisse für die Seminare Diagnostik Grundlagen zu bewerten. Denn neben der vom Verfasser dieser Arbeit konstruierten, unbenoteten Abschlussklausur des Seminars mussten alle Studierenden am gleichen bzw. nächsten Tag an der verpflichtenden, von Prof. S. Buch konstruierten, benoteten Klausur der Vorlesung Pädagogische Diagnostik teilnehmen. Im Grundlagenseminar und der Vorlesung standen relativ viele ähnliche Themen an, wenn auch in unterschiedlicher Akzentuierung und Intensität. Die Korrelation der beiden gänzlich unabhängig voneinander konstruierten Diagnostikklausuren von $r = .72$ ($N=56$) belegt eine erstaunliche Übereinstimmungsvalidität der Messungen und besagt, dass derjenige, der in der unbenoteten Abschlussklausur der hier untersuchten Seminare gut abgeschnitten hat, auch in der benoteten Diagnostik-Vorlesung-Klausur gute Ergebnisse erzielte. Obwohl somit alle Studierenden zeitnah auch eine benotete Klausur ähnlicher Thematik erwarteten und sich entsprechend darauf vorbereiten mussten, gelang es den Studierenden nicht, die anfänglichen Defizite des unbenoteten Quiz in der unbenoteten Abschlussklausur völlig auszugleichen. Auch deshalb werte ich die hier gefundenen Leistungsvorteile der benoteten Quiz als ziemlich konservativ und näher an der unteren Grenze dessen, was man normalerweise erwarten sollte.

Benotung im Studium erfordert rationales Lernmanagement

Die Einführung der Bachelor-Studiengänge verschärfte die Verschulung des Studiums mit der Konsequenz, dass die hier untersuchten Studierenden in jedem Seminar notenrelevante Leistungen erbringen mussten, die dann in bestimmter Form in ihr Abschlusszeugnis eingehen. Es wäre im Hinblick auf die Sicherung eines möglichst guten Studienabschlusses suboptimal, unter diesen Bedingungen an sich förderliche, aber nicht zwingend notenrelevante Lernangebote besonders intensiv zu nutzen, weil ein derartiges Engagement Zeit verschlingt, die der prüfungsrelevanten Vorbereitung oder der Erholung bzw. Freizeit dann fehlt. **So betrachtet sprechen die Ergebnisse für eine systemimmanente**

Rationalität der Studierenden, bei unbenoteten Quiz weniger Lerneinsatz zu zeigen. Sie sprechen aber nicht zwingend dagegen, auch unbenotete Tests mit Rückmeldungen anzubieten, da diese, wie ja nachgewiesen wurde, ebenfalls langfristiges Behalten fördern, wenn auch in geringerem Umfang. Die unbenoteten Onlinequiz mit vielfältigen Rückmeldungen waren als verpflichtende Hausaufgaben und nicht als optionale Lernhilfen angesetzt worden. Freiwillige Onlinequiz würden sich theoretisch lohnen, wenn ihre Bearbeitung bessere benotete Testergebnisse nach sich zöge. Darüber hinaus wäre es vorteilhaft, die Überlegenheit von Onlinequiz gegenüber einfacheren Lernformen nachzuweisen, was durchaus auch misslingen kann (z.B.: Jacobs 2008).

Benotung als akzeptables Mittel zu Leistungssteigerung ?

Unter pädagogischer Perspektive werden benotete Tests im Schulsystem hauptsächlich aus diagnostischen Erwägungen zum Zwecke der Leistungsbewertung- bzw. Zertifizierung eingesetzt. Zugleich entfachen sie jedoch ein Lernbemühen, wie es kaum durch sonstige pädagogische Maßnahmen zu erzielen ist. Dadurch sichern sie quasi unabhängig vom Interesse am Lerngegenstand ein hohes Ausmaß erforderlicher Lernzeit ab und bewirken mehr Lernerfolg als didaktisch durchaus vernünftige und interessante Lernangebote ohne direkte Notenkonsequenz.

Manch einer mag sich angesichts der Ergebnisse dieser Studie auf den Standpunkt stellen, nur noch notenrelevante Leistungen abzuverlangen, was stellenweise ja bereits geschieht. Der hohe diagnostische Aufwand ließe sich reduzieren, in dem lediglich Zufallsstichproben der zu erbringenden Leistungen tatsächlich benotet werden würden. Wie Daniel & Broida (2004) aufzeigten, lassen sich auch benotete Online-Quiz ohne externe Kontrollen konzipieren, die in ihrer Lernwirksamkeit streng kontrollierten benoteten Quiz im Seminar vergleichbar sind. Vermutlich reicht bereits die Mindesterwartung einer möglichen Benotung aus, langfristig ein hinreichendes Bedrohungspotenzial aufrecht zu erhalten. **So würde sich freilich das Bildungssystem als eine auf Zwang ausgerichtete Kontrollinstanz qualifizieren, die eine selbstverantwortliche Schwerpunktbildung und Vertiefung des Wissens nahezu verunmöglicht, stattdessen Anpassung und Unterwerfung verlangt und nur noch wenig mit dem Begriff des Studiums gemein hat.** Zudem wäre bei allgegenwärtig permanenter Benotung irgendwann der Punkt erreicht, ab dem Ermüdung, Resignation oder Überforderung das Auspressen weiterer Lernbemühungen verhindern würden. Es kommt offenbar auf das rechte Maß an, benotete Leistungen einzufordern und es muss vor allem hinreichend viele Lernphasen bzw. Seminare ohne jeden Prüfungscharakter geben.

Bonussystem für unbenotete Quiz als Alternative ?

Um die Leistungsbereitschaft für unbenotete Quiz zu steigern, bietet sich ein leistungsabhängiges Bonussystem an, bei dem der Studierende entsprechend seiner Quizleistung Punkte erhält, welche die Gesamtpunktzahl bei der Seminarendnote nur erhöhen können. Während Benotung häufig bedroht, da man seine Seminarendnote verschlechtern kann, fokussiert der Bonus ausschließlich auf ein positives Hinzugewinnen. Dann könnte man auch damit rechnen, fast alle Studierende nähmen freiwillig diese Tests in Anspruch (Kibble 2007). Theoretische Überlegungen, wie auch die bisherigen empirischen Befunde (z.B.: Hautau et. al. 2006a, 2006b, Sansgiry et al. 2006, Oliver & Williams 2005) lassen Leistungsvorteile eines Bonussystems gegenüber einer Testung ohne jede extrinsische Konsequenz erwarten. Jedenfalls sehe hier eine ernsthafte Alternative zur konventionellen Quizbenotung sowie einen schwachen Hoffnungsschimmer, gute Leistungen mit weniger Stress zu erreichen.

Eine etwas umfangreichere Version dieses Beitrags findet der interessierte Leser unter.

http://www.phil.uni-sb.de/~jakobs/wwwartikel/extrinsische_motivierung/notendruck/notenanreiz.html

Literatur:

- Azorlosa, J. L & Renner, C. H. (2006). The Effect of Announced Quizzes on Exam Performance. *Journal of Instructional Psychology*, Vol. 33, No. 4. 278-283.
- Bangert-Drowns, R., Kulik, J., & Kulik, C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, 85(2), 89-99.
- Carroll, J. B. (1963). A model of school learning. *Teacher College Record*, 64, 723-733.
- Daniel, D. B. & Broida, J. (2004). Using Web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, 31, 207-208.
- Elikai, F. & Baker, J. (1988). Empirical Evidence on the Effectiveness of Quizzes as a Motivational Technique. *Issues in Accounting Education*, Fall88, Vol. 3 Issue 2, p248, 7p
- Harris, V. W. & Sherman, J. A. (1974). Homework assignments, consequences, and classroom performance in social studies and mathematics. *Journal of Applied Behavior Analysis*, 7, 505-519.
- Hautau, B., Turner, H.C., Carroll, E., Jaspers, K., Krohn, K., Megan, P., Williams, R.L. (2006a). Differential Daily Writing Conditions and Performance on Major Multiple-Choice Exams. *Journal of Behavior Education*, 15:171-181
- Hautau, B., Turner, H. C., Carroll, E., Jaspers, K., Parker, M. · Krohn, K., Williams, R. L. (2006b). Differential Daily Writing Contingencies and Performance on Major Multiple-Choice Exams. *Journal of Behavior Education*, 15: 259-276.
- Jacobs, B. (1981). *Angst in der Prüfung - Beiträge zu einer kognitiven Theorie der Angstentstehung in Prüfungssituationen*. R. G. Fischer-Verlag. Frankfurt/Main.
- Jacobs, B. (1996-2005). *Situative, aktuelle Prüfungsangst (SPA)*.
<http://www.phil.uni-sb.de/FR/Medienzentrum/verweise/psych/wwwfrage/spajava2.html>
- Jacobs, B. (2003a). Die Wirkung einer Probeklausur auf Lernleistung und aktuelle Angst in einer echten Klausur.
<http://www.phil.uni-sb.de/~jakobs/wwwartikel/probeklausur/index.htm>
- Jacobs, B. (2007). Geld und Noten als extrinsische Motivatoren zur Verbesserung kognitiver Leistungen.
 URN: urn:nbn:de:bsz:291-psydok-9644
 URL: <http://psydok.sulb.uni-saarland.de/volltexte/2007/964/>
- Jacobs, B. (2008). Gezieltes Studieren gelöster Aufgaben als alternative Übungsmethode zu Testen mit Feedback.
 URN: urn:nbn:de:bsz:291-psydok-15597
 URL: <http://psydok.sulb.uni-saarland.de/volltexte/2008/1559/>
<http://www.phil.uni-sb.de/~jakobs/wwwartikel/testread/testread2.html>
- Jacobs, B. (2009): Geld zur Steigerung der Testmotivation- und Leistung in einer unbenoteten Abschlussklausur ?
 URN: urn:nbn:de:bsz:291-psydok-23600
 URL: <http://psydok.sulb.uni-saarland.de/volltexte/2009/2360/>
http://www.phil.uni-sb.de/~jakobs/wwwartikel/extrinsische_motivierung/cash4quiz.html
- Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Advances in Physiological Education* 31, 253 - 260.
- Narloch, R., Garbin, C. P. & Turnage, K.D. (2006). Benefits of Prelecture Quizzes. *Teaching of Psychology*. 33 (2) , 109-112.
- Oliver, R. & Williams, R. L. (2005). Direct and Indirect Effects of Completion Versus Accuracy Contingencies on Practice-Exam and Actual-Exam Performance. *Journal of Behavioral Education*, Vol. 14, No. 2, June 2005 (C 2005), pp. 141-152.
- Ryan, C. S. & Hemmes, N. S. (2005). Effects of the contingency for homework submission on homework submission and quiz performance in a college course. *Journal of Applied Behavior Analysis*. 38, 79-88.
- Sansgiry, S.S., Chanda, S., Lemke, Th. & Szilagyi, J.E (2006). Effect of Incentives on Student Performance on Milemarker Examinations. *American Journal of Pharmaceutical Education* 2006; 70 (5) Article 103.

 Die Forschung wurde unterstützt aus Studiengebühren der Universität des Saarlandes