

**Bernhard Jacobs**, Fachrichtung Erziehungswissenschaften der Universität des Saarlandes.  
 Email: b.jacobs@mx.uni-saarland.de  
 Version: 6.1.2011

## Musterlösungen durcharbeiten als Alternative zu Testen mit Feedback - Eine Replikationsstudie

### Abstract

Ziel der Untersuchung war eine Replikation der Studie von Jacobs 2010c, die überwiegend den Nachweis erbrachte, im realen Universitätsbetrieb erziele das Testen mit unmittelbarem Aufgaben spezifischen Feedback einen ähnlich guten Lernerfolg wie die Bearbeitung von Musterlösungen. Teilweise erbrachte die Studie aber auch unerwartete Befunde sowie offene Fragen und regte so eine erneute Überprüfung mit etlichen Änderungen an. Gegenüber Jacobs 2010c wurde das Retentionsintervall reduziert, nur solche Onlineübungen herangezogen, die aufgrund früherer Studien bessere Lernleistungen erbrachten als gar keine Maßnahme, bei den Testungen ein erforderliches Minimalleistungskriterium eingeführt und eine milde Kontrolle zur Bearbeitung von Musterlösungen eingesetzt. Alle durchgeführten Analysen führten zu hoch vergleichbaren Lernerfolgsmaßen beider Übungsmethoden bei etwas geringeren Lernzeiten für das Durcharbeiten korrekter Antworten. Studierende halten die Testung insgesamt für die bessere und lernwirksamere Übungsvariante, würden bei freier Übungswahl aber eher Musterlösungen durcharbeiten.

**Schlagerworte:** Testen, Feedback, Quiz, Musterlösung, Üben, Studieren, Lernförderung

## Einleitung, Problemstellung und Zielsetzung

Nachdem hinreichende Befunde die positive Wirkung des Testens, sowie den über eine Testung hinaus gehenden Nutzen des Feedbacks auf das Behalten nachgewiesen haben, konzentriert sich die neuere Forschung zunehmend auf die Frage, ob das Testen auch dem Studieren korrekt gelöster Testfragen überlegen ist, also die aktive Fragebeantwortung mit anschließender Rückmeldung der korrekten Antwort mehr Lerngewinn verspricht als das Bemühen, direkt eine Musterlösung nachzuvollziehen. Die Unterschiede zwischen beiden Übungsmethoden fallen bisherigen empirischen Studien zufolge in der Regel nicht hoch aus, tendieren aber insbesondere in der Laborforschung zu einer leichten Bevorteilung der Tests, was sich vornehmlich bei Verwendung von Short Answer- bzw. Free Recall-Aufgabentypen nachweisen ließ (siehe einige Literaturangaben dazu bei Jacobs 2010c). In der psychologischen Memory-Forschung, etwa beim Vokabellernen, wird der Testvorteil meist mit unterschiedlichen Anforderungen an das Gedächtnis erklärt, wobei insbesondere das Testen einen anstrengenden Abruf aus dem Gedächtnis einfordert, dadurch Zugang und Zugriff auf die relevante Information verbessert und so ein stabileres Behalten bewirkt (retrieval hypothesis). Die Lernwirksamkeit des Feedbacks basiert vornehmlich auf der Korrektur der Fehler.

Der Vergleich Testen vs. Studieren der korrekten Antworten wirft zugleich einige didaktische Grundsatzprobleme auf, etwa die, wie umfangreich und ausdifferenziert der Lehrstoff vermittelt bzw. vom Lernenden dabei aktiv mit vollzogen werden sollte, ob eine explizite Testung tatsächlich notwendig ist oder implizit auch vom Lerner übernommen werden kann, ob bzw. wann von der Präsentation des Lehrstoffs zu einer Testung übergegangen werden soll und welche Rolle hierbei interaktive Lernformen (Testen mit Feedback) übernehmen könnten (Jacobs 2003).

Im Vergleich zur Laborforschung (z.B.: LaPorte & Voss 1975, Cull 2000, Roediger & Karpicke 2006) werden in der Schulpraxis höhere pädagogische Anforderungen gestellt. So fällt der Stoffumfang meist höher aus, der Abstand zwischen der Lernaneignung und Übung liegt zeitlich weiter auseinander, das Lehrzielniveau geht deutlich über Faktenwissen hinaus, man erwartet zumindest einen minimalen Transfer des Wissens sowie eine gewisse zeitliche Nachhaltigkeit des Gelernten. Unter der Berücksichtigung dieser Anforderungen fand Jacobs (2008a, 2010c) bisher überwiegend

vergleichbare Lerneffekte für beide Übungsmethoden und nur in einem Fall einen Testvorteil. Andererseits berichten einige Autoren auch im realen Universitätssetting von besseren Ergebnissen einer Testung mit Feedback gegenüber dem Studieren korrekter Antworten (z.B.: Clifton, 2005, McDaniel, Anderson, Derbish & Morrisette, 2007; Anderson, Weywadt & McDaniel, 2009; Larsen, Butler & Roediger, 2009). Vorliegende Studie sollte unter anderem prüfen, ob sich die bisherigen eigenen Befunde replizieren lassen. Dabei entspricht die angestrengte Untersuchung keiner stringenten, identischen Replikation. Zum Teil enthält sie etliche Abweichungen, welche an den entsprechenden Stellen angesprochen werden.

## Untersuchungsansatz

An der Untersuchung nahmen Studierende des Lehramts teil, die ein vom Verfasser dieser Arbeit geleitetes Proseminar zur Pädagogischen Diagnostik im SS 2010 belegten. Das erste Seminar fand montags von 12.30 -14 Uhr, das zweite Seminar dienstags von 10-11.30 statt. Beide Seminare wurden analog Jacobs 2010c vom selben Dozenten nach identischem Seminarplan unterrichtet und hatten den gleichen Lehrstoff mit denselben Lehrmaterialien zu bearbeiten. Da die experimentellen Übungen als normale Übungsbestandteile im Seminar integriert waren, wurden die Studierenden nicht darüber informiert, dass sie an einem Experiment teilnahmen und erhielten erst nach Abschluss aller Erhebungen Informationen über den Sinn und Zweck der Untersuchung.

Der Seminarleiter gestaltete alle für die experimentellen Übungen relevanten Seminarsitzungen überwiegend dozierend und bemühte sich, einen möglichst gleichartigen Unterricht in beiden untersuchten Seminaren zu halten. Die jeweiligen Sitzungen wurden durch umfangreiche Powerpoint-Folien unterstützt, die den Studierenden unmittelbar nach der Seminarsitzung verfügbar gemacht wurden und als hinreichende Lektüre für die Veranstaltung bezeichnet wurden.

Zum einen wurde die Fragestellung auf der Basis eines vorexperimentellen Versuchsplans im Seminar überprüft, zum anderen bearbeiteten die Studierenden im Rahmen eines Feldexperimentes zwei Onlineübungen individuell zu Hause. Tabelle 1 verdeutlicht den Zeitpunkt der Übungen sowie der Kriteriumstests, welche die langfristigen Auswirkungen aller Übungen auf das Behalten erfassen sollte. Alle hier berichteten Zeitangaben beziehen sich auf das Seminar 1.

**Tabelle 1:** Zeitpunkte für die Übungen und die Lernerfolgstests

10.5.2010	20.5.2010	14.6.2010	1.7.2010	12.7.2010
Übung im Seminar	Online Übung 1	Kriterium 1 Behaltenstest	Online Übung 2	Kriterium 2 Behaltenstest

Alle Übungen beziehen sich auf den Lehrstoff einer Seminarsitzung von 90 Minuten. Übungs- und Lernerfolgsaufgaben sind nicht identisch, sondern ähnlich (parallele Aufgaben). Der Behaltenstest bzw. das Kriterium 1 erfasst die Lehrziele aus der Seminarübung „Aufgabenformen“ und der ersten Onlineübung "Skalen und Stichproben", der Behaltenstest 2 bezieht sich auf die Lehrziele der zweiten Onlineübung („Aufgabenanalyse“). Mithin ergeben sich für die Behaltensprüfungen der einzelnen Übungen unterschiedliche Retentionsintervalle:

Übung im Seminar : 35 Tage  
 Online-Übung 1 : 25 Tage  
 Online-Übung 2 : 11 Tage

Der Zeitabstand zwischen Übung und Überprüfung des Lernerfolgs betrug bei Jacobs 2008b eine Woche, bei Jacobs 2010c mindestens vier, stellenweise sogar 10 Wochen. So wünschenswert ein langfristiges Behalten auch sein mag, mit zunehmender Zeit nimmt das Vergessen seinen Lauf und umso schwieriger wird es, mögliche Unterschiede von Übungsmethoden nachzuweisen. Die meisten mir bekannten Studien legen Behaltensintervalle von höchstens einer Woche zugrunde. Ein weiterer Grund, diesmal die Retentionsintervalle zu verkürzen, lag in der Notwendigkeit des Über-

raschungsmomentes sowie der Organisation der Lehrveranstaltung. Es könnte sich mittlerweile herumgesprochen haben, dass gegen Ende des Seminars unverhofft Testungen anstehen. Aber die Lernerfolgsmaße sollten für die Studierenden ohne Vorankündigung, also völlig unerwartet zum Einsatz kommen.

Alle Lernerfolgstests fanden im Seminar unter kontrollierten Bedingungen statt. Im Gegensatz zu Jacobs 2010c wurde auf die besondere pädagogische Motivierungsmaßnahme (siehe Jacobs, 2009b) zur Bearbeitung der unbenoteten Kriteriumstests verzichtet. Der Seminarleiter ging jedoch kurz vor dem Austeilen der Kriteriumstests auf die das Lernen fördernde Funktion von Testungen ein, propagierte die wissenschaftlichen Belege für eine Behaltensstabilisierung durch Tests und versprach, individuelle Rückmeldung zum Prozentsatz der korrekten Lösung zu geben. Die Antworten der Studierenden im Kriteriumstest wurden in eine dem Testheft analoge Computerfassung übertragen und automatisch ausgewertet.

Einige Studierende gaben das Seminar im Verlauf des Semesters auf. Wenige Studierende beantworteten nicht alle Fragen zu den subjektiven Angaben. Sehr wenige Studierende versäumten das rechtzeitige Bearbeiten der Übungen, waren bei den Lernerfolgstests aber zugegen. Die so in Kauf zu nehmenden, marginalen Datenausfälle führen zu geringfügig unterschiedlichen Probandenanzahlen für die einzelnen Vergleiche. Wo immer möglich, wurden alle verfügbaren Daten zur Datenanalyse herangezogen. Die experimentellen Vergleiche mit den Kriteriumstests beinhalten natürlich nur solche Studierende, die auch an den Übungen teilgenommen hatten. Für die statistische Prüfung des Lernerfolgs wird ein Signifikanzniveau von 10% zugrunde gelegt, für sonstige Variablen ein Signifikanzniveau von 5%.

## Studie 1: Testen bzw. Musterlösungen in einer Seminarsitzung

Die Überprüfung der Hypothese im Seminar entspricht einer identischen Replikation zur Studie 1 von Jacobs 2010c, weswegen sich die Ausführungen auf das absolut Notwendige beschränken. Der Seminarleiter setzte überraschend eine spezielle Übung via Computer und Beamer zum Inhalt der vorausgehenden Sitzung "Aufgabentypen und Aufgabenformen" an. Seminar 2 bearbeitete das Quiz Aufgabenformen in Form der bei Jacobs 2010c beschriebenen Testung durch Handzeichen mit unmittelbar anschließendem Feedback (KOR, KCR und elaboriertes Feedback) nach jeder Aufgabe, während dem Seminar 1 dieselben Fragen als Musterlösungen präsentiert wurden. Eine Musterlösung umfasste die Fragestellung, die korrekte(n) Antwort(en) und die Erklärung analog dem elaborierten Feedback. Die Auswirkungen der beiden Methoden wurden 5 Wochen später in einem unangekündigten unbenoteten Behaltenstest 1 (siehe Tabelle 1) geprüft. Versuchsplantechnisch liegt diesem Vorgehen der vorexperimentelle statistische Gruppenvergleich ohne Vormessung zugrunde. Tabelle 2 belegt nahezu identische Ergebnisse beider Seminare hinsichtlich des Abiturnotendurchschnitts und annähernd vergleichbare Werte bzgl. Alter und Geschlecht.

**Tabelle 2:** Vergleichbarkeit der Seminare 1 und 2 (Seminar 1 N = 22-26; Seminar 2: N= 26-29)

	Seminar 1		Seminar 2		
	M	s	M	s	
Abiturnotendurchschnitt:	2.33	0.73	2.33	0.47	ns
Alter:	24.7	5.3	22.7	2.3	ns
Geschlecht:	61% w		66% w		ns

Die hoch vergleichbare Leistungsfähigkeit beider Seminare im Bezug auf die Themen der pädagogischen Diagnostik wird auch durch die Ergebnisse der sonstigen Kriteriumstests untermauert (siehe weiter unten). Beide Seminare erzielten sowohl im Behaltenstest Skala und Stichproben (59,7 vs. 59,1) wie auch im Kriteriumstest Aufgabenanalyse (60,1 vs. 60,0) fast identische Mittelwerte des Prozentsatzes der korrekten Lösungen. Alle Ergebnisse sprechen trotz des "formal schwachen Versuchsplans" für eine sehr hohe interne Validität.

## Ergebnisse

### *Bearbeitungszeiten in den Seminaren*

Die Bearbeitungszeiten fielen für die Testung mit 34 Minuten etwas länger aus als für die Musterlösungen mit 30 Minuten. Ein Teil dieses Zeitunterschieds dürfte auf die anfänglichen Erklärungen der besonderen Antwortprozedur durch Karten und Handzeichen beim Testen zurückgehen. Pragmatisch gesehen ist der Zeitunterschied zu vernachlässigen. Im Gegensatz zu den Onlineübungen in Studie 2, welche eine individuelle Übungsbearbeitung und damit auch individuelle Bearbeitungszeiten erlauben, muss der Seminarleiter im Seminar bei seinen Zeitvorgaben möglichst alle Studierenden im Blick haben.

### *Behaltenswerte im Kriteriumstest*

Der nach 5 Wochen angesetzte Lernerfolgstest erzielte eine recht schwache Konsistenz von  $\alpha = .53$ , was teilweise auch an der Heterogenität der Aufgaben liegen mag und nicht zwingend eine schwache Zuverlässigkeit signalisieren muss. So korreliert der Kriteriumstest Aufgabenformen mit den Kriterientests Skala  $r = .54$  und Aufgabenanalyse  $r = .46$  in ansprechender Höhe. Jeweils ein Studierender aus beiden Bedingungen wurde wegen extrem schwacher Testergebnisse (Extremwerte nach Tukeys Boxplot) aus der Analyse ausgeschlossen.

**Tabelle 3** : Prozentsatz korrekter Lösungen im Lernerfolgstest Aufgabenformen, SS 10

Seminar	Exp. Bedingung	M	s	N	t	p <sub>z</sub>
2	Testen	63.3	9.4	27	0.25	.81
1	Musterlösung	62.5	13.4	25		

Wie aus Tabelle 3 hervorgeht, lässt sich kein signifikanter Unterschied zwischen beiden Treatments feststellen. Für den langfristigen Lernerfolg spielte es demnach keine Rolle, ob die Studierenden im Seminar offen aktiv die Quizfragen beantworteten und erst nach Ihrer Antwort eine informative Rückmeldung (Knowledge of Correct Response + gegebenenfalls elaboriertes Feedback) erhielten oder ob sie die bereits korrekt beantworteten Quizfragen einschließlich des elaborierten Feedbacks lediglich durchzulesen hatten, wobei sie sich um ein Verständnis bemühen sollten. Beide Instruktionmethoden führten, genauso wie bei Jacobs 2010c, zu hoch vergleichbaren Behaltenswerten im Sinne eines nahen Anforderungstransfers.

## Studie 2: Testen bzw. Musterlösungen als verpflichtende Online-Übungen

### *Testen mit Feedback*

Bei Jacobs 2010c wurde besonderer Wert darauf gelegt, ein Übungsangebot ohne jeglichen Notendruck anzubieten. Die bisherigen Erfahrungen mit gänzlich unbenoteten, wenn auch verpflichtenden, Onlinequiz ließen aber die Vermutung aufkommen, manche Studierende würden diese Art der Testung nicht hinreichend ernst nehmen. Einige, glücklicherweise sehr wenige Studierende, lieferten sogar Quizergebnisse im Bereich von Zufallsantworten ab. Um derartige Reaktionen möglichst gänzlich zu unterbinden, wurde diesmal ein Mindestkriterium der Quizleistung von 50% eingeführt. Dies wurde im Seminar mehrmals angekündigt und stand in einer Anweisung unmittelbar vor Bearbeitung des Quiz. Das untersuchte Seminar wird zwar offiziell nicht benotet. Dennoch müssen bestimmte Leistungen erbracht werden, um eine erfolgreiche Teilnahme beim Prüfungsamt bestätigt zu bekommen. Vor Beginn des Quiz sah der Studierende unter anderem folgende Textpassage: „Sie sollten wenigstens 50% der Aufgaben korrekt lösen können.“ Die Maßnahme scheint die er-

hoffte Wirkung erzielt zu haben, da in beiden Quiz jeweils nur ein Studierender dieses Minimalkriterium verfehlte.

Auch in dieser Studie kommt mangels hinreichender Studentenzahl keine No-Treatment-Kontrollgruppe zum Einsatz. Jedoch belegen frühere experimentelle Untersuchungen mit den hier eingesetzten Onlinequiz, dass diese gegenüber keiner Maßnahme einen Behaltensvorteil erwarten lassen. Die erste Onlineübung bei Jacobs 2010c 'Statistische Grundlagen der pädagogischen Diagnostik' wurde hier durch die Übung 'Skala und Stichproben' ersetzt, weil sich in einem früheren Seminar die Lernwirksamkeit des entsprechenden Quiz gegenüber keiner Maßnahme statistisch belegen ließ (Jacobs 2010b). Auch die zweite hier eingesetzte Onlineübung Aufgabenanalyse konnte ihre Lernwirksamkeit früher bereits nachweisen (Jacobs 2009a bzw. 2010c gegenüber Musterlösungen) und sollte darüber hinaus wegen des unerwarteten Ergebnisses bei Jacobs 2010c einer erneuten Überprüfung unterzogen werden. Die angenommene Lernwirksamkeit der hier eingesetzten Onlinequiz scheint somit in dieser Untersuchung empirisch gut begründet. Sollte das Studieren von Musterlösungen einen ähnlichen Lernerfolg nach sich ziehen, so gibt es begründeten Anlass, auch dieser Methode eine Lernwirksamkeit gegenüber keiner Maßnahme zuzusprechen.

Alle Aufgaben eines Quiz waren auf einer HTML-Seite untergebracht. Vor den Aufgaben erhielten die Studierende nähere Tipps, wie sie Aufgaben bearbeiten sollten.

1. Versuchen Sie jeweils zunächst, die Aufgabe selbständig zu lösen.
  2. Klicken Sie dann auf den Button "Aufgabe bestätigen". Im Rückmeldefenster sehen Sie die Aufgabenerklärung. Es ist sehr wichtig, sich die Erklärungen verständlich zu machen.
  3. Klicken Sie anschließend auf den Button "Korrekte Lösung". Dadurch wird die Aufgabe automatisch korrekt beantwortet. Überprüfen Sie die Aufgabe im Hinblick darauf, ob Sie die Lösung nachvollziehen können.
- Bei verbleibenden Unklarheiten haben Sie jederzeit die Möglichkeit durch Anklicken auf "Aufgabe bestätigen" die Erklärungen wieder einzusehen.

Der Studierende musste jede Aufgabe beantworten bzw. bestätigen. Unmittelbar nach der Aufgabenbestätigung hatte der Übende Zugriff auf folgende Rückmeldungen

Knowledge of Response (richtig oder falsch)  
 Knowledge of Correct Response (die korrekte Antwort)  
 elaboriertes Feedback, [nur dann wenn es als förderlich angesehen wurde].

Erst nach Beantwortung aller Aufgaben konnte der Studierende die Übung zu beenden und damit die notwendige Zertifizierung für die Übungsteilnahme erhalten. Im Anschluss an die Übungsbeendigung folgte dann noch das Feedback „Prozentsatz der korrekt gelösten Aufgaben.“

### *Musterlösungen*

In der Studie von Jacobs 2010c erzielte das Testen bei der Übung Aufgabenanalyse entgegen der Erwartung einen höheren Lernerfolg als das Studieren der Musterlösungen. Speziell bei der Übung Aufgabenanalyse war der Versuch unternommen worden, die Übungsbedingungen einer Musterlösungs-Variante zu optimieren, in dem bestimmte, als unnötig erachtete Testinformationen, etwa Distraktoren bei einer MC-Aufgabe, herausgenommen wurden, um unnötig erscheinenden Text einzusparen. Obwohl stellenweise ähnliche Optimierungsversuche bei Jacobs 2008a keine schwächeren Ergebnisse produzierten als das Onlinequiz, wurde diesmal die einfachere Originalversion verwendet. Sie ist mit einer Testung identisch, umfasst somit auch die Fragestellung und unterscheidet sich lediglich durch die unmittelbare Präsentation der korrekten Antwort etwa in Form einer sichtbaren Markierung der korrekten Antwort bei einer MC-Aufgabe einschließlich der elaborierten Rückmeldungen.

Bei Jacobs 2010c wurde wohlwollend unterstellt, Studierende würden die dargebotenen Musterlösungen in konstruktivem Sinne sorgfältig und vollständig bearbeiten. Sie gelangten auf eine WWW-Seite, die alle Musterlösungen umfasste und am Ende lediglich einen Button bereit hielt, der die Beendigung der gesamten Übung ermöglichte. Diese großzügige Freiheit führte bei einigen Studierenden zu sehr kurzen Bearbeitungszeiten, welche den Verdacht aufkommen ließen, diese hätten sich sehr oberflächlich mit den Musterlösungen befasst. Bei den Onlinequiz musste demgegenüber jede Aufgabenbearbeitung bestätigt werden, was dann die Rückmeldung "richtig/falsch", die korrekte Antwort, sowie gegebenenfalls weitere Erklärungen nach sich zog. Um den Lernenden auch bei den Musterlösungen zu einer aufmerksamen und vollständigen Bearbeitung zu bewegen, musste der Lernende hier analog dem Quizteilnehmer jede Musterlösung bestätigen. Diese Bestätigung durch jeweiliges Anklicken auf den Button "Bearbeitung der Musterlösung bestätigen" verwandelte diesen in den Button "ok" und war notwendig, um die Übung beenden zu können. Die Übungsbeendigung schließlich war Voraussetzung für die Zertifizierung der Übung als Seminarleistung. Eine derartige "Kontrolle" bei den Musterlösungen fehlte bei Jacobs 2010c, kam aber beispielsweise bei Hodge (2009), Clifton (2005) und McDaniel et al. (2007) zur Anwendung. Diese Maßnahme garantiert natürlich keine tiefe Verarbeitung, mag aber eine gewisse Disziplinierung bewirken, die Musterlösungen wenigstens zur Kenntnis zu nehmen. Zudem sollten Empfehlungen die vernünftige Bearbeitung der Musterlösungen anregen:

1. Lesen Sie die Musterlösungen aufmerksam durch und versuchen Sie, sich die Lösungen verständlich zu machen. - Scrollen Sie erst danach zu den weiteren Erklärungen.
2. Zur eigenen Kontrolle könnten Sie die vorgegebenen Fragen mit geschlossenen Augen und eigenen Worten beantworten bzw. begründen.
3. Suchen Sie nach Möglichkeiten, die dargebotenen Informationen mit eigenen Erfahrungen zu verbinden.

## Versuchsdurchführung

Zu Beginn des Semesters wurden die Studierenden innerhalb jedes der beiden Seminare nach Zufall auf die Gruppen A und B zugeteilt, bei der statistischen Analyse und Auswertung dann die A- bzw. B- Gruppen der einzelnen Seminare zu einer gemeinsamen A- bzw. B-Gruppe zusammengefasst. Im Gegensatz zu Jacobs 2010c erhielten die Studierenden keinerlei Information darüber, zu welcher Gruppe sie gehörten, weil sie diese Information hier nicht zwingend benötigten. Tabelle 4 bestätigt sehr hohe Ähnlichkeiten beider Gruppen hinsichtlich Abiturnotendurchschnitt, Alter und Geschlecht.

**Tabelle 4:** Vergleichbarkeit der Gruppen A (N=23-25) und B (N=29)

	A		B		
	M	s	M	s	
Abiturnotendurchschnitt:	2.30	0.60	2.33	0.56	ns
Alter:	23.5	4.1	23.6	5.7	ns
Geschlecht:	64%w		62%w		ns
Kriterium Aufgaben:	61.8	17.5	61.2	11.7	ns

Die hoch vergleichbaren Ergebnisse beider Gruppen im Lernerfolgstest Aufgabenformen unterstreichen auch vergleichbare Leistungsfähigkeiten beider Gruppen im Hinblick auf die Lehrziele des Faches pädagogische Diagnostik. Die Zuweisung der Studierenden zu den Gruppen A und B entspricht somit einer sehr erfolgreichen Randomisierung und liefert die Grundlage für den in Tabelle 5 dargestellten experimentellen Versuchsplan.

**Tabelle 5 : Versuchsplan der Onlineübungen**

Online Übung 1 20.5.10		Kriterium 1 14.6.2010	Online Übung 2 1.7.2010		Kriterium 2 12.7.2010
<b>R A:</b>	<b>Testen 1</b>	<b>O</b>	<b>Musterlösung 2</b>	<b>O</b>	
<b>R B:</b>	<b>Musterlösung 1</b>	<b>O</b>	<b>Testen 2</b>	<b>O</b>	

Als erste Übung absolvierte Gruppe A das Quiz zum Thema Skala und Stichproben, während Gruppe B zur selben Zeit die Musterlösungen bearbeitete. Bei der zweiten Übung, ca. 5 Wochen später, stand das Thema Aufgabenanalyse an. Nun wurden die Übungsbedingungen für beide Gruppen getauscht. Gruppe B bearbeitete die Übung als Quiz mit unmittelbarem Feedback und Gruppe A als Studieren der Musterlösungen. Unmittelbar vor und nach jeder Übung beantworteten die Studierenden anonym ein paar Fragen, die einige wichtige Informationen zur Interpretation der Ergebnisse beisteuern sowie eine Bewertung der Übungsmethoden aus Sicht der Studierenden ermöglichen sollten. Ca. dreieinhalb Wochen nach der ersten Online-Übung und ca. 10 Tage nach der zweiten Onlineübung nahmen alle Studierenden an einer unangekündigten, unbenoteten Testung im Seminar teil, welche für jede Übung 10 parallele Aufgaben enthielt, die zusammengefasst das jeweilige Behaltenskriterium einer bestimmten Übung messen sollten.

Im Seminarplan, der via Internet jederzeit eingesehen werden konnte, waren die Termine der Onlineübungen exakt festgelegt worden. Abweichend von Jacobs 2010c wurden die experimentellen Übungen im Seminarplan nicht als Übungen, sondern stets als Quiz angekündigt. Die Teilnahme an den Onlineübungen war verpflichtend und wurde elektronisch überprüft. Zu keinem Zeitpunkt wussten die Studierenden, zu welcher Gruppe sie gehörten. Vor einem Übungstermin mussten alle Studierenden also ein Quiz erwarten, da sie erst beim Starten der Übung davon Kenntnis erhielten, welche Übungsform sie zu bearbeiten hatten. Lediglich das Thema der Übung war bekannt. Es bezog sich auf die Inhalte der vorausgehenden Seminarsitzung.

### *Zusammenfassung der wichtigsten Rahmenbedingungen für die Onlinequiz*

Da die Lernwirksamkeit von Tests bzw. Übungen von etlichen Konstellationen abhängt, welche teilweise über die Eigenschaften der speziellen Übungen hinausgehen (siehe z.B. Jacobs 2010a), gewährt Tabelle 6 einen Überblick der wichtigsten Faktoren.

**Tabelle 6 : Überblick der wichtigsten sonstigen Bedingungen für die Onlinequiz.**

Übungsankündigung:	klar festgelegt und lange bekannt, stets als Quiz angekündigt, aber tatsächlich: Quiz oder Musterlösung
Übungsteilnahme	verpflichtend, notwendig für die Zertifizierung des Seminars
Übungsgrundlage	vorausgehende Seminarsitzung und Powerpoint-Folie des Dozenten.
Übungstermin	jeweils 1 bis 3 Tage nach übungsrelevanter Seminarsitzung
Übungsrelevante Lehrziele	transparent. (Übungsbeispiele in der Powerpoint-folie)
Übungsablauf	Zeit unbeschränkte, extern unkontrollierte Onlineübung von zu Hause via Internet
Übungsbewertung	nur für Quiz: Mindestleistung 50% korrekte Antworten erforderlich
Aufgabenbearbeitung	bei Quiz: jede Aufgabe musste aktiv bestätigt werden, bei Musterlösung: jede Musterlösung musste aktiv bestätigt werden.
Ergebnisrückmeldung	bei Quiz: Prozentsatz korrekter Lösungen bei Musterlösung: keine, da nicht möglich
Lernerfolg/Kriterium	unangekündigt, unbenotet, parallele Aufgaben: kontrollierte Bearbeitung im Seminar

## Aufgabenbeispiel für Quiz, Musterlösung und Lernerfolgstests

Quizaufgaben und Musterlösungen sollten exakt die gleichen Informationen enthalten und sich lediglich dadurch unterscheiden, dass beim Quiz zunächst eine Antwort erforderlich ist, bevor die Rückmeldung folgt und die Musterlösung bereits die korrekte Antwort und alle Rückmeldungen enthält. Beide Übungsformen hatten zum Ziel, das Verständnis zu fördern, weswegen die Aufgaben des Lernerfolgstests die Lehrziele der Übungsaufgaben mit nicht identischen, sondern ähnlichen Aufgaben erfassen sollten. Hierzu soll aus beiden Übungen je ein Beispiel gezeigt werden.

### Beispiel aus der Übung Skala und Stichproben

Die Übung selbst ist übersichtlicher gestaltet und musste hier aus Platzgründen eng zusammengefasst präsentiert werden.

Auf die Darstellung der **Quizaufgabe 7** wurde aus Platzgründen verzichtet.

Die Quizaufgabe entspricht der Musterlösung 7, nur muss der Studierende die Antworten selbst geben und dann auf den Button „Aufgabe bestätigen“ klicken. Er erhält dann KOR, sowie die elaborierte Rückmeldung, die hier bei der Musterlösung direkt sichtbar ist. Durch Anklicken auf korrekte Lösung wird die Aufgabe automatisch korrekt beantwortet und man sieht dann wie bei der Musterlösung 7 die richtigen Antworten.

Bei der Musterlösung muss der Studierende hier wie bei allen Musterlösungen den Button „Bearbeitung der Musterlösung bestätigen“ anklicken.

#### Musterlösung 7

Im folgenden werden Skalenwerte (und gelegentlich die Variable) angegeben, die gemessen werden sollen. Sie sollen das höchstmögliche Skalenniveau angeben, welches eine Skala mit den vorgegebenen Skalenwerten annehmen kann. Es kann auch sein, dass sich die Werte nicht zu einer Skala zuordnen lassen (=keine Skala).

	Skalenwerte	a nominal	b ordinal	c intervall	d rational	e keine Skala
1.)	Intelligenztestwerte: 70, 110, 140	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
2.)	rot, grün, blau, gelb	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3.)	1 Meter, 30 Zentimeter, 2 km	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
4.)	Leistung: Gold-, Silber-, Bronzemedaille	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5.)	Deutsche, Juden, Ausländer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
6.)	ziemlich geil, sehr geil, turbooberaffengeil	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7.)	Wirbeltiere, Weichtiere, Affen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
8.)	Unterschicht, Oberschicht, Mittelschicht	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Bearbeitung der Musterlösung bestätigen !

Minimale Voraussetzung für jede Skala sind:

Jedem Element kann und muss genau ein Skalenwert zugeordnet werden. Die möglichen Kategorien einer Skala müssen sich ausschließen. Werden keine darübergehenden Anforderungen gestellt, so liegt eine **Nominalskala** vor. Werden die Bedingungen nicht erfüllt, so liegt

**keine Skala** vor, z.B. Es gibt deutsche Juden. (=Kategorien schließen sich nicht aus, sondern überschneiden sich). Der Gorilla Bimbo ist ein Affe, ein Säugetier und ein Wirbeltier. (Ein Skalenwert schließt andere Skalenwerte ein)

**Ordinalskala:** Die Werte können in eine Rangordnung gebracht werden. Die Abstände zwischen den Skalenwerten sind aber unbekannt bzw. unklar.

**Intervallskala:** mindestens Ordinalskala und darüber hinaus: Die Abstände zwischen den Skalenwerten sind gleich (äquidistant).

**Rationalskala, auch Verhältnisskala:** Mindestens Intervallskala und darüber hinaus. Empirisch sinnvoller absoluter Nullpunkt, z.B. 0 kg.

#### Lernerfolgstest Aufgabe 7

Im folgenden werden einige Skalenwerte (und gelegentlich die Variable) angegeben, die gemessen werden sollen. Sie sollen das höchstmögliche Skalenniveau angeben, welches eine Skala mit den vorgegebenen Skalenwerten annehmen kann. Es kann auch sein, dass sich die Werte nicht zu einer Skala zuordnen lassen (=keine Skala).

	Skalenwerte	a nominal	b ordinal	c intervall	d rational	e keine Skala
1.)	10°, 20° und 25° [Grad Celsius]	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2.)	nie, gelegentlich, häufig, sehr häufig	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3.)	5 kg, 100 Gramm, 2 Tonnen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4.)	Erster, Zweiter, Letzter	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5.)	Nationalität: Italiener, Schweizer, Sizilianer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6.)	Interesse: kaum, mäßig, hoch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7.)	Säugetiere, Vögel, Waale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8.)	Uniabschluss: Promotion, Master, Bachelor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



## Beispiel aus der Onlineübung Aufgabenanalyse

<p><b>Quiz Aufgabe 10</b></p> <p>Welcher Funktion folgt die Beziehung zwischen Aufgabenschwierigkeit und Trennschärfe einer Aufgabe am ehesten ?</p> <p>1.) <input type="radio"/> linear          2.) <input type="radio"/> u-förmig          3.) <input type="radio"/> umgekehrt u-förmig          4.) <input type="radio"/> exponentiell</p> <div style="display: flex; justify-content: space-between; margin-top: 10px;"> <span>Aufgabe bestätigen</span> <span>Korrekte Lösung?</span> </div>	<p><b>Lernfolgstest Aufgabe 10</b></p> <p>Wie lässt sich die Beziehung zwischen Aufgabenschwierigkeit und Trennschärfe einer Aufgabe am besten beschreiben          Mit wachsender Aufgabenschwierigkeit...</p> <p>1.) <input type="radio"/> wächst auch die Trennschärfe kontinuierlich an          2.) <input type="radio"/> sinkt die Trennschärfe zunächst und wächst dann an.          3.) <input type="radio"/> wächst die Trennschärfe zunächst an und nimmt dann ab.          4.) <input type="radio"/> wächst die Trennschärfe erst langsam, dann schneller an.</p> <div style="text-align: center; margin-top: 10px;"> <span>Aufgabe bestätigen</span> </div>
<p><b>Musterlösung 10</b></p> <p>Welcher Funktion folgt die Beziehung zwischen Aufgabenschwierigkeit und Trennschärfe einer Aufgabe am ehesten ?</p> <p>1.) <input type="radio"/> linear          2.) <input type="radio"/> u-förmig          3.) <input checked="" type="radio"/> umgekehrt u-förmig          4.) <input type="radio"/> exponentiell</p> <p>Die Trennschärpen folgen dieser Funktion nicht im deterministischen Sinne.</p> <p>Jedoch gilt allgemein. Mittlere Aufgabenschwierigkeiten erzielen die höchsten Trennschärpen. Je mehr die Aufgabenschwierigkeiten von 50% abweichen, desto geringer fallen im Durchschnitt die Trennschärpen aus.</p> <div style="text-align: center; margin-top: 10px;"> <span>Bearbeitung der Musterlösung bestätigen !</span> </div>	

Beim Quiz kann der Lernende die korrekte Lösung erst dann erhalten, wenn er die Aufgabe beantwortet bzw. bestätigt hat. Nach der Aufgabenbestätigung sieht der Lernende in einem Rückmeldefenster genau dieselben Informationen, die bei der Musterlösung direkt noch angefügt sind.

Bei der Musterlösung muss der Lernende den Button „Bearbeitung der Musterlösung bestätigen“. Dieser Button verwandelt sich dann in den Button „ok“

## Ergebnisse

### *Leistungsergebnisse in den Quiz*

Tabelle 7 zeigt die deskriptiven Ergebnisse der beiden Quiz aus vorliegender Untersuchung. Nach Tukey's Boxplot gab es trotz großer Leistungsvarianz keinerlei Ausreißer, so dass alle verfügbare Studierenden in der Auswertung eingingen. Die Reliabilitäten der Quiz und der Lernerfolgsmaße sind nur zum Teil zufrieden stellend. Insbesondere die schwache Konsistenz beim Kriterium Aufgabenanalyse von  $\alpha = .54$  fiel schwächer aus als bisheriger Erfahrung nach ( $\alpha = .71$  bei Jacobs 2010c) zu erwarten war. Der Wert spricht jedoch nicht zwingend für eine völlig unzureichende Zuverlässigkeit, sondern eher für einige heterogene Anforderungen, da die Retestreliabilität ganz ordentlich ausfällt und beide Kriterien immerhin .61 miteinander korrelieren.

**Tabelle 7:** Deskriptive Ergebnisse zu den experimentellen Quiz

	Anzahl Aufgaben	Reli $\alpha$	M	s	N
Testen 1	15	.72	69.4	15.4	25
Testen 2	15	.74	79.5	16.4	28
Kriterium 1	10	.75	58.8	19.6	54
Kriterium 2	10	.54	60.0	17.3	52
Parallelretest Testen 1 - Kriterium 1	r = .60 (N=25)				
Parallelretest Testen 2 - Kriterium 2	r = .54 (N=27)				
Korrelation zwischen Kriterium 1 und 2	r = .61 (N=52)				

Einige Daten geben Anlass zu der Vermutung, bei den Online-Tests sei es nicht immer mit rechten Dingen zugegangen. Dies gilt insbesondere für die recht hohen Erfolgsquoten im Quiz 2, die bisherigen Erfahrungen zufolge im Mittel ca. 10 % zu hoch ausfielen. Dort konnten drei "exzellente Leistungen mit 100% korrekter Antworten" sowie zwei Studierende identifiziert werden, die im kontrollierten Kriteriumstest einen Leistungseinbruch gegenüber dem Onlinequiz von mehr als 40 % aufwiesen, was die normale Vergessensrate deutlich übersteigt. Man muss vermutlich davon ausgehen, die Einführung einer Mindestleistung im Quiz habe einige wenige –ziemlich evident – zwei Studierende dazu bewegt, ihre Quizwerte irgendwie zu schönen. Dadurch wird zwar deren Messung diagnostisch entwertet, die Lernwirkung des Testens aber noch keineswegs zwingend aufgehoben. Ohne jeden Zweifel müssen wichtige offizielle Onlineprüfungen unter kontrollierten Bedingungen durchgeführt werden. Aber auch unkontrollierte Onlinequiz lassen sich durch bestimmte Maßnahmen (z.B. zufällige Auswahl von Aufgaben, sehr enges Zeitfenster für die Bearbeitungszulassung, Zeitbegrenzung der Aufgabenbearbeitung usw.) besser kontrollieren. Derartige Restriktionen waren hier aber nicht vorgesehen, da die Maßnahme ja nicht primär auf Leistungsbeurteilung, sondern eher auf ein möglichst natürliches Üben abzielte, wobei das geforderte Minimal-kriterium lediglich die Ernsthaftigkeit des Übens unterstützen sollte.

### *Bearbeitungszeiten der Onlineübungen*

Da das Durcharbeiten bereits korrekt beantworteter Quiz weniger Zeit erfordert als die eigenständige Beantwortung von Quiz, war mit kürzeren Bearbeitungszeiten für Musterlösungen gerechnet worden, zumal die Untersuchungen von Jacobs 2007a und 2008a sowie die Übung "Statistische Grundlagen der pädagogischen Diagnostik" aus Jacobs 2010c diese Hypothese übereinstimmend empirisch eindeutig belegten. Lediglich die Zeiten zur Übung Aufgabenanalyse bei Jacobs 2010c ergaben, der Tendenz nach zwar ebenfalls Daten in erwarteter Richtung, aber keine signifikant geringere Bearbeitungszeiten für die Musterlösungen, was mit einer gesonderten Intervention des Seminarleiters zu erklären versucht wurde, welche diesmal unterblieb. Infolgedessen wird nun auch bei der Übung 2 (Aufgabenanalyse) eine kürzere Bearbeitungszeit für die Musterlösungen erwartet. Zunächst wurden aus Übung 1 mittels Tukeys Boxplot ein Ausreißer mit überlangen Zeiten aus der Bedingung Musterlösungen, sowie bei der zweiten Übung zwei Studierende mit überlangen Übungszeiten aus der Bedingung Testen mit Feedback ausgeschlossen. Mit den so modifizierten Daten wurde ein t-Test für unabhängige Stichproben berechnet. Zusätzlich wurde der Wilcoxentest unter Einbezug der Ausreißer durchgeführt. Die Signifikanztestung wurde wegen der klaren Hypothesenrichtung einseitig formuliert.

Wie aus Tabelle 8 hervorgeht, entsprechen alle Zeiten der Erwartung und ergeben kürzere Bearbeitungszeiten der Musterlösungen gegenüber den Testungen. Der Zeitunterschied liegt im mittleren Effektstärkebereich. Die Bearbeitung korrekt beantworteter Quizfragen erfordert ca. 70 bis 80% der für die eigene Quizbeantwortung erforderlichen Lernzeit.

**Tabelle 8:** Übungszeiten in Minuten für beide Übungen

	M	s	N	t	pe	Wilcoxon z-Wert	pe
Testen 1	30.9	13.3	25	1.75	.043	1.4	0.079
Musterlösung 1	25.1	10.2	27				
Testen 2	18.7	7,7	26	2.52	.008	2.82	0.005
Musterlösung 2	13.3	7.7	25				

*Ergebnisse im Lernerfolgstest*

Die bisherigen eigenen Studien sprechen überwiegend für eine vergleichbare Übungseffektivität beider Übungsvarianten, was die Ergebnisse in Tabelle 9 übereinstimmend bestätigen. Denn in beiden Onlineübungen erzielten Quiz und Musterlösung hoch vergleichbare Leistungswerte im jeweiligen Kriteriumstest.

**Tabelle 9:** Prozentsatz korrekter Lösungen im Lernerfolgstest (Kriterium)

Gruppe	Exp. Be- dingung	M	s	N	t	pz
A	Testen 1	59.7	20.4	25	0.34	.72
B	Musterlösung 1	57.9	19.1	29		
B	Testen 2	59.1	17.5	27	0.07	.94
A	Musterlösung 2	59.5	17.2	23		

Insbesondere die der Erwartung voll entsprechenden, nahezu identischen Ergebnisse der Übungsvarianten in Übung 2 geben Anlass zu der Vermutung, das unerwartete Ergebnis von Jacobs 2010c sei auf Sonderfaktoren der Bedingung Musterlösung zurück zu führen, die nicht der Normalität entsprechen. Wegen der günstigeren Bearbeitungszeiten erweisen sich die Musterlösungen demnach bei gleicher Lerneffektivität als lerneffizienter.

## Subjektive Daten

Die Erhebung subjektiver Daten unmittelbar vor (Angstmaße) sowie nach der Übung (sonstige Maße) sollte die Vergleichbarkeit der Gruppen im Hinblick auf wichtige Lern- und Motivationsvariablen prüfen. Die in Tabelle 10 dargestellten Ergebnisse entsprechen bis auf wenige Ausnahmen in hohem Maße den früheren Befunden von Jacobs (2010c).

Die Analyse belegt über beide Übungen hinweg eine sehr hohe Vergleichbarkeit im Hinblick auf die angegebene Vorbereitungszeit einer Übung sowie hohe Ähnlichkeit im Hinblick auf subjektiven Angaben zur Güte der Vorbereitung. Beide Übungsvarianten wurden mit der gleichen Anstrengungsbereitschaft bearbeitet und ein gutes Ergebnis in einem potentiellen Test zum Thema der Übung hat für beide Gruppen eine ähnlich hohe Bedeutung. Aus subjektiver Sicht liegen demnach eine vergleichbare Übungsvorbereitung sowie Lern- und aktuelle Übungsmotivation für beide Übungstypen zugrunde. Es gelang auch hier nicht, Unterschiede in der Akzeptanz, Bewertung oder Benotung der Übung zu finden, wobei zu beachten ist, dass die Studierenden offenbar nicht die Übungsalternative in Erwägung zogen. Im Gegensatz zu den früheren Ergebnissen erlebten die Studierenden das Quiz 2 Aufgabenanalyse anstrengender als die Studierenden, welche die bereits korrekt gelösten Aufgaben zur Aufgabenanalyse bearbeiteten ( $t(49)=2.7$   $p=0.01$  zweiseitig). Dies war jedoch beim Quiz 1 nicht der Fall.

**Tabelle 10** Ergebnisse zu subjektiven Einschätzungen**a) Übung 1:** Skala und Stichproben, Pädagogische Diagnostik SS 10

N(Quiz) 18-25 N(Muster)=23-29

	Reli $\alpha$	Quiz M s		Musterlösung M s	
aktuelle Prüfungsangst SPA	.94	27.2	10.2	26.0	10.1
Fearthermometer FT	.71	3.1	2.7	3.2	2.4
Quizvorbereitung	.73	3.7	0.9	3.9	0.8
Zeit für Quizvorbereitung(min)		44.2	28.1	49.3	29.6
Bedeutsamkeit	.75	3.8	0.8	3.8	0.9
Anstrengungsbereitschaft	.68	4.2	0.7	4.0	0.6
Übung ist anstrengend		3.2	0.8	2.9	1.0
Akzeptanz der Übung	.83	3.4	0.5	3.8	0.7
Note für Übung		2.5	0.7	2.2	0.7

**b) Übung 2** :Aufgabenanalyse, Pädagogische Diagnostik SS 10

N(Quiz) 26-27 N(Muster)=24-25; \* signifikanter Unterschied

	Reli $\alpha$	Quiz M s		Musterlösung M s	
aktuelle Prüfungsangst SPA	.97	26.9	11.1	22.3	12.5
Fearthermometer FT	.72	3.0	2.3	3.2	2.7
Quizvorbereitung	.69	3.7	0.9	3.9	0.6
Zeit für Quizvorbereitung(min)		37.5	30.0	44.4	28.7
Bedeutsamkeit	.77	3.8	0.8	3.7	0.9
Anstrengungsbereitschaft	.85	3.7	0.8	3.8	0.8
Übung ist anstrengend*		3.0	0.9	2.3	1.0
Akzeptanz der Übung	.78	3.8	0.5	3.5	0.7
Note für Übung		2.4	0.8	2.6	0.9

**Direkter Vergleich beider Übungsmethoden aus Sicht der Studierenden**

Nachdem alle Studierenden im Verlauf des Seminars Erfahrungen mit Tests und Musterlösungen gemacht hatten, sollten sie die beiden Übungsmethoden im direkten Kontrast gegeneinander einschätzen. Dazu wurde in der letzten Seminarsitzung, am 12.7.2010 in Seminar 2 nach dem Kriteriumstest 2 und im Seminar 1 vor dem Kriteriumstest 2 der bereits bei Jacobs 2010c eingesetzte Kurzfragebogen im Seminar zur Beantwortung vorgelegt. Aufgabe der Studierenden war es, bestimmte Fragen bzw. Bewertungen komparativ für Quiz oder Musterlösung zu entscheiden, wobei als dritte Alternative auch eine Unentschiedenheit möglich war. Auf die Frage "Welche Übungsmethode bewirkt ein besseres Verständnis" konnte der Studierende das Quiz, die Musterlösung oder "weder Quiz noch Musterlösung" ankreuzen. Die Antwortkodierung, Quiz = 1, Musterlösung=0, "weder Quiz noch Musterlösung" = 0.5 erlaubt eine direkte Interpretation der Bevorzugung einer bestimmten Methode. Je näher der Wert bei 1 liegt, desto eher sprechen die Ergebnisse für eine Bevorzugung von Quiz und damit Testen mit Feedback.

Die bei Jacobs 2010c verwandte Skala Präferenz der Übung, die sich aus dem Mittelwert der ersten 6 Items aus Tabelle 12 zusammensetzt, welche inhaltlich auf eine bessere Bewertung der Methode im Hinblick auf Lernwirksamkeit, Verständnis und Interesse sowie eine Bevorzugung der entsprechenden Methode bei Wahlfreiheit hindeuten, erweist sich hier jedoch trotz eines  $\alpha$  von .76 als nicht so homogen wie erwartet und lässt bereits auf der Basis einer Faktorenanalyse mit Varimax-Rotation zwei Faktoren erkennen. Eine schiefwinklige Rotation ergibt dann schließlich eindeutig zwei Faktoren, die inhaltlich am ehesten als Einschätzung der pädagogischen Qualität (Item 1..4,

Ladungen: .52-.90.) sowie eigene Bevorzugung der Methode (Item 5 und 6; Ladungen .90-.95) zu deuten sind. Zwar ergaben die Daten von Jacobs 2010c eine einfaktorielle Lösung, wobei das Item "ist insgesamt als die bessere Übungsmethode einzuschätzen" dort die höchste Ladung (.84) und das Item "würde ich bei Wahlfreiheit selbst auswählen" die geringste Ladung (.68) erbrachte. Im Nachhinein betrachtet, war die von Jacobs 2010c verwendete Bezeichnung Präferenz jedoch unglücklich und missverständlich gewählt. Theoretisch lassen sich die Aspekte „Qualität der Übungsmethoden“ und „Bevorzugung einer Übungsmethode“ hier jedenfalls klar voneinander trennen.

Bei Jacobs 2010c erhielten die Studierenden einen Präferenzwert von .78, der deutlich höher ausfällt als hier (.58). Der mögliche Grund für die unterschiedlichen Ergebnisse lässt sich aus Tabelle 11 sowie einer detaillierten Analyse der Items in Tabelle 12 erahnen.

**Tabelle 11:** Ergebnisse zur Übungsbewertung sowie Binomialtest ( $p < .5$  vs.  $p > 0.5$ ) zweiseitig

	$\alpha$	M	s	p
Qualität der Übungsmethode	.72	<b>.66</b>	.30	.001
Bevorzugung der Übungsmethode	.74	<b>.41</b>	.37	.164
Präferenz analog Jacobs 2010c	.76	<b>.58</b>	.28	.007

Aus Tabelle 11 und 12 geht hervor, dass die Studierenden die pädagogische Qualität des Testens deutlich höher einschätzen als die von Musterlösungen, aber eher zu einer Bevorzugung von Musterlösungen tendieren. Tabelle 12 stellt die Mittelwerte und die prozentualen Häufigkeiten der Entscheidungen "Quiz, Musterlösung oder weder Quiz noch Musterlösung" für alle Items der Präferenz analog Jacobs 2010c (1..6) sowie für einige weitere Items dar. Die meisten Studierenden schätzen Quiz mit Feedback zwar als die bessere Übungsmethode ein und halten sie für lernwirksamer, würden bei Wahlfreiheit aber eher Musterlösungen auswählen. Das war bei Jacobs 2010c anders. Dort wollten 66 % aller Studierenden bei Wahlfreiheit eher die Quiz auswählen und es waren auch deutlich mehr Studierende der Ansicht, man solle öfter Quiz als Musterlösungen anbieten. [Bevorzugung der Übungsmethode bei Jacobs 2010c :  $M = .75$ ,  $s = .32$ ]. Aus einer guten pädagogischen Qualitätsbewertung der Quiz lässt sich hier somit keine direkte Bevorzugung folgern.

**Tabelle 12:** Prozentuale Entscheidungen für Quiz oder Musterlösung,  
Mittelwert nach Kodierung und Ergebnis des Binomialtests (1 vs. 0) ( $N = 52$ )

Item Nr.	Item reihenfolge	Welche Übungsmethode.....	Quiz %	Musterlösung %	Mittelwert Kodierung	Sign. p
1	3	ist insgesamt als die bessere Übungsmethode einzuschätzen	67	21	<b>.73</b>	.001
2	8	weckt bei der Bearbeitung ein höheres Interesse	46	17	<b>.64</b>	.014
3	10	schätze ich als lernwirksamer ein	67	21	<b>.74</b>	.001
4	12	bewirkt ein besseres Verständnis	38	35	<b>.52</b>	ns
5	11	sollte man öfter als Übungen anbieten	35	35	<b>.50</b>	ns
6	7	<b>würde ich bei Wahlfreiheit selbst auswählen</b>	23	58	<b>.33</b>	.008
7	1	erfordert eine größere geistige Anstrengung	85	6	<b>.89</b>	.000
8	4	ermöglicht eine bessere Einschätzung des eigenen Wissens	73	13	<b>.79</b>	.000
9	5	verleitet eher zu oberflächlichem Arbeiten	4	81	<b>.12</b>	.000
10	6	erfordert weniger Lernzeit und Lernaufwand	2	85	<b>.90</b>	.000

Möglicherweise ist die unterschiedliche Einschätzung in beiden Studien durch das hier gesetzte Leistungskriterium für die Quiz zu begründen. Jacobs (2010c) legte besonderen Wert darauf, jede Form der Quizbenotung auszuschließen, um jeglichen Stress zu vermeiden. Er machte mit aller

Deutlichkeit klar, das Quizergebnis werde nicht gewertet. Um unzumutbare schwache Leistungen zu verhindern, wurde hier jedoch eine Mindestleistung im Quiz gefordert, die ihren angestrebten Zweck auch weitgehend erfüllte. Die Studierenden wussten, dass sie mindestens 50% aller Quizfragen richtig beantworten mussten.

Angesichts der massiven Leistungsnachweise, mit denen insbesondere Studierende der neuen Studiengänge konfrontiert werden, bleibt es nachvollziehbar, freiwillig keine weiteren Leistungsnachweise herbeizusehen. Wenn man die Wahl hat, eine weniger anstrengende Übungsmethode auswählen, fällt es nicht schwer, diese Chance auch nutzen, selbst wenn deren Lernwirksamkeit schwächer eingeschätzt wird. Schließlich geht es auf einer Universität nicht um Wissen an sich oder Bildung um ihrer selbst willen, sondern darum, die notwendigen Anforderungen für einen Studienabschluss zu erfüllen. Die restlichen Items stimmen in hohem Maße mit den Befunden von Jacobs 2010c überein und bestätigen die Einschätzung aus Sicht der Studierenden, Testen sei aufwändiger, anstrengender, ließe mehr Verarbeitungstiefe erwarten und liefere eine bessere Einschätzung des eigenen Wissens.

Zusätzlich sollten die Studierenden den pädagogischen Nutzen beider Methoden auf einer Ratingskala zwischen 0 = unbrauchbar und 10 = hervorragend sowie die Qualität jeder Übungsmethode anhand der üblichen Notenskala von 1 bis 6 einschätzen.

**Tabelle 13:** Ratings zum pädagogischen Nutzen und der Qualität der Übungsmethoden, t-test für abhängige Stichproben (N=51-52) und Effektstärke.

	Quiz		Musterlösung		t	d
	M	s	M	s		
Pädagogischer Nutzen (0-10):	7.3	1.4	5.7	2.4	4.6	0.83
Qualität via Note (1-6):	2.2	0.6	2.8	1.0	-4.0	0.71

Die Ergebnisse in Tabelle 13 bestätigen immer noch recht überzeugend die eindeutig bessere Bewertung der Quiz gegenüber den Musterlösungen, die Größe des Unterschieds fällt aber augenscheinlich geringer aus als bei Jacobs 2010c. Der Testvorteil bei der Qualitätseinschätzung via Note erweist sich mit  $d = .71$  sogar als signifikant niedriger als bei Jacobs 2010c mit  $d = 1.53$  ( $z = -2.57$ ,  $p < .05$ , Korrelationsdifferenzentest nach Umwandlung der Effektstärken in Korrelationen).

### Exkurs: Erlauben Quiz tatsächlich eine bessere Einschätzung des eigenen Wissens?

Nach einer beide Übungsmethoden vergleichenden subjektiven Einschätzung der Studierenden erlaubt das Quiz in beiden Studien eindeutig eine bessere Einschätzung des eigenen Wissens als die Vorlage von Musterlösungen (siehe für diese Studie: Tabelle 12). Diese Wertung erscheint natürlich dann trivial, wenn die objektiven Quizergebnisse direkt mitgeteilt werden, was in der Regel bei Onlinetests möglich ist. Die Studierenden wurden deshalb unmittelbar nach ihrer Übung, unter Quizbedingung bewusst noch vor der Mitteilung des Gesamtergebnisses in Form des Prozentsatzes der korrekten Lösungen, danach gefragt, wie hoch sie den Prozentsatz ihrer korrekten Lösungen einschätzen: "Wenn Sie jetzt an einem Quiz zu "Skalen und Stichproben" teilnehmen würden, wie viel Prozent der Aufgaben würden Sie vermutlich korrekt beantworten?". Wegen der Anonymität der subjektiven Angaben kann die Analyse leider nur Mittelwertsergebnisse berücksichtigen. Zwar darf aufgrund empirischer Untersuchungen angenommen werden, bei den Quiz korreliere die subjektive Schätzung und das objektive Quizergebnis relativ hoch miteinander, für Musterlösungen ist dieser Zusammenhang allerdings unbekannt.

Tabelle 14 stellt den objektiven Ergebnissen die subjektiven Schätzungen der beiden Übungsmethoden gegenüber. Wegen der Zufallszuweisung zu den Übungsmethoden sind die tatsächlichen Quizergebnisse auch gute statistische Schätzungen für die Studierenden unter Bedingung Musterlösungen. Den Mittelwertsergebnissen der Tabelle 14 zufolge, kann allein aufgrund der Übungsbearbeitung kaum von einer besseren Gesamtleistungseinschätzung der Quizteilnehmer die Rede sein.

Obwohl die Quizteilnehmer während der Quizbearbeitung nach jeder Aufgabenbestätigung die Rückmeldung "Knowledge of Response: richtig / falsch: sowie den Prozentsatz korrekter Lösungen bei Mehrfachantworten erhielten, erzielten sie im Mittel keine genaueren Schätzungen.

**Tabelle 14:** Mittelwerte des tatsächlichen Prozentsatzes korrekter Lösungen für die Quizgruppe und der subjektiven Schätzungen unter beiden Übungsbedingungen

		objektiv Quiz	subjektiv Quiz    Musterlösung	
Jacobs 2010c	Übung 1	72.3	75.6	68.2
	Übung 2	67.1	62.0	68.6
hier	Übung 1	69.4	70.3	77.7
	Übung 2	79.5	72.9	77.4

Insgesamt reflektieren die Ergebnisse Methoden übergreifend einen ziemlich realistischen Eindruck, gewichten meiner Meinung nach aber teilweise etwas zu wenig den Lernzuwachs durch die Übung. Die Daten lassen kaum Kompetenzillusionen bei der Bearbeitung korrekt gelöster Aufgaben erkennen, wie in der Literatur häufiger für reine Studiermethoden im Sinne reiner Informationsaufnahme berichtet wird. Der Studierende hat offenbar auch bei der Präsentation von Musterlösungen gewisse Vorstellungen davon, wie gut er diese selbst gelöst hätte und welche Themen er vielleicht noch nacharbeiten müsste. Durch die Mitteilung des Prozentsatzes der korrekten Lösungen erhält aber schließlich nur der Quizteilnehmer die Gewissheit, ob seine Schätzung zutrifft. Insofern liefern Quiz durch die Ergebnismitteilung eine objektive Basis für die Selbsteinschätzung der Gesamtleistung und durch die unmittelbare Rückmeldung KOR bei jeder Aufgabe ein sehr spezifisches objektives Feedback, was die Studierenden offenbar bei Ihrer Bewertung deutlich haben einfließen lassen.

## Exkurs: Aktuelle Angst in einem P/F- Quiz

Auch die unmittelbar vor der Übung erhobenen aktuellen Angstmittelwerte beider Übungsvarianten erbrachten keine signifikanten Unterschiede. Dieses, auf den ersten Blick gänzlich unerwartete Ergebnis mag zum Teil daran liegen, dass die Studierenden ein Quiz erwarteten und erst beim Starten der Übung darüber informiert wurden, welche Übung (Quiz oder Musterlösung) sie tatsächlich zu bearbeiten hatten. Obwohl im Vorspann der Übung mit Musterlösungen, unmittelbar vor dem Angsttest, darauf hingewiesen wurde, es gehe in der Übung darum, sich bereits korrekt gelöste Aufgaben verständlich zu machen, kann man davon ausgehen, viele Studierende hätten diese Information nicht nachhaltig aufgenommen und irgendeine Form der Testung erwartet. Insofern dürfte der aktuelle Angstvergleich Quiz gegen Musterlösung irreführend sein und wird nicht weiter verfolgt. Auf jeden Fall sagt er nichts darüber aus, welche Werte zu erwarten wären, wenn die Studierenden im Vorfeld den Übungstyp hinreichend antizipieren könnten.

Da hier (SS 10) im Quiz eine Mindestleistung von 50% verlangt wurde, die einer Pass/fail-Benotung [ $\geq 50\%$  = bestanden;  $< 50\%$  = nicht bestanden] ähnelt, müsste man höhere Angstwerte vermuten als in gänzlich unbenoteten Quiz. Die Ergebnisse deuten zumindest in diese Richtung und sollen durch ein paar Vergleichsdaten aus früheren Studien des gleichen Seminars bei den gleichen Quiz belegt werden, obwohl diese Form des Vergleichens über verschiedene Semester hinweg methodisch natürlich sehr angreifbar ist (Jacobs 2009c).

**Tabelle 15:** Aktuelle Angst im Quiz 1 (Skala und Stichproben) sowie Quiz 2 (Aufgabenanalyse) für einige Semester.

Quiz 1 Skala und Stichproben

	benotet		unbenotet	
	M	s	M	s
WS 08	25.6	10.7	17.2	8.2
WS 09			21.9	10.5
SS 10	<b>27.2</b>	10.2		

Quiz 2 Aufgabenanalyse

WS 08	15.1	7.9
WS 09	16.2	11.2
SS 10	<b>26.9</b>	11.1

Die Benotung im WS 08 entsprach der üblichen mehrstufigen Benotung (1-5).

Die Benotung im SS 10 entsprach einer Pass/fail Benotung (50% Kriterium:  $\geq 50\%$  bestanden, ansonsten nicht bestanden)

Die Anzahl der Studierenden beträgt jeweils ca.  $N = 25$ .

Die aktuelle Angst wurde mit dem Angstfragebogen SPA von Jacobs (1981) erhoben

Im WS 08 konnte auf experimentellem Weg nachgewiesen werden, dass konventionell benotete Quiz signifikant mehr aktuelle Angst unmittelbar vor der Testung auslösen als gänzlich unbenotete Quiz (Jacobs 2009a). Für das Quiz Skala und Stichproben des WS08 beträgt die Effektstärke des signifikanten Unterschieds  $d=.88$ . Man erkennt aus der Tabelle 15, dass die Angstwerte der P/F-Benotung denen einer klassischen mehrstufigen Quizbenotung recht vergleichbar sind und die Angst unter P/F-Benotung grob geschätzt ca. eine halbe Standardabweichung höher ausfällt als die gänzlich unbenoteter Quiz. In gewisser Hinsicht kann dies als schwacher empirischer Hinweis dafür gewertet werden, durch eine Notensystemveränderung vom mehrstufigen Notensystem auf P/F könne man keineswegs die aktuelle Angst vor Prüfungen reduzieren (siehe Jacobs 2010 d).

## Zusammenfassung und Diskussion

### Hauptergebnis

Insgesamt kann von einer recht gelungenen Replikation gesprochen werden, die einige Unklarheiten bei Jacobs 2010c beheben konnte, aber auch neue Erkenntnisse beisteuerte. **Alle Ergebnisse bestätigen die vergleichbare Lerneffektivität von Testen mit Rückmeldung und dem Durcharbeiten von Musterlösungen. Damit hat sich das Studieren korrekt gelöster Aufgaben als eine dem Testen mindestens ebenbürtige Übungsmethode erwiesen, zumal die Lerneffizienz in Form kürzerer Lernzeiten sogar besser ausfällt.** Im Unterschied zu den meisten empirischen Studien zum Vergleich „Testen vs. gezieltes Studieren“, die in der Regel identische Aufgaben in der Übung und im Lernerfolgstest beinhalten, erforderte der Kriteriumstest hier einen nahen Anforderungstransfer. Anders als beim Lernen von Vokabeln, wo erfolgreicher Retrieval in einer Übung das langfristige Behalten deutlich stärkt, erschöpfen sich anspruchsvollere Lehrziele nicht in der Memorierung der korrekten Antwort. Auch die Lernwirksamkeit des Feedbacks schwindet mit wachsender Aufgabenkomplexität (Kluger & DeNisi, 1996). Entscheidend für den Lernerfolg dürfte letztlich das ernsthafte Bemühen des Lernenden sein, die ihm verfügbare Übungsgrundlage selbst gründlich durcharbeiten. Hierzu kann explizites Testen mit Feedback etliche Anregung geben, es ist aber nicht zwingend erforderlich, sondern stellt eine unter mehreren Übungsmöglichkeiten dar.

### Einschränkungen

Diese Interpretation gilt freilich nur unter den hier gesetzten Restriktionen und darf nicht unreflektiert auf alle möglichen Übungssituationen extrapoliert werden. Im Wesentlichen wird vorausgesetzt, die Studierenden würden gleich gut vorbereitet zur Übung antreten und diese auch mit annähernd gleicher Lern- bzw. Testmotivation bearbeiten. Diese Bedingungen lassen sich vor allem in solchen Lernsituationen relativ einfach realisieren, die eine gewisse Kontrolle beim Lernerwerb und



den Übungen ermöglichen, z.B. unmittelbar nach einer Instruktion oder Stillarbeit im Klassenverband oder in einer unerwarteten Übung im Seminar. Die Bedingungen sind umso schwieriger zu erfüllen, je mehr Eigeninitiative und Selbststeuerung dem Lernenden abverlangt werden, etwa nach dem Lesen eines Kapitels im Lehrbuch oder der Vor- bzw. Nacharbeitung einer Seminarsitzung ohne erkennbare direkte Überprüfung. Hier dürften beide Übungsmethoden nur bei solchen Studierenden in ähnlicher Weise fruchten, welche die Übungen mit hohem Interesse bzw. hinreichender Lernmotivation in Angriff nehmen.

Ansonsten könnten die beiden Übungsmethoden bereits Einfluss auf die vorausgesetzte Lernaneignung ausüben, welche in der Übung gefestigt werden soll. Im Gegensatz zur Präsentation von Musterlösungen, kann das Testen durch Androhen leistungsabhängiger Konsequenzen wesentlich einfacher als Disziplinierungsvehikel genutzt werden, um im Vorfeld eine bessere Lernaneignung zu erzwingen sowie während der Übung eine höhere Lernmotivation anzuregen (Jacobs 2009a). In vorliegender Studie täuschte der Seminarleiter die Studierenden unter der Bedingung Musterlösung, indem er ihnen einfach ein Quiz ankündigte, was letztlich eine zur Testgruppe vergleichbar gute Vorbereitung bewirkte. Eigenen Angaben zufolge hatten die Studierenden beide Übungen auch mit gleicher Anstrengungsbereitschaft bearbeitet und den Wert guter Leistungen vergleichbar hoch eingeschätzt, obwohl nur die Quizgruppe harte Konsequenzen zu erwarten hatte. Die Onlineübungen waren im Übrigen so konzipiert, dass sie ein allzu lasches oder ablenkendes Arbeitsverhalten durch milde Kontrollen zu verhindern suchten. So konnte während der Übungszeit nur die Übung auf dem Computer genutzt werden und jede Aufgabe bzw. jede Musterlösung musste einzeln bestätigt werden. Obwohl die Kontrolle von Musterlösungen beschränkt ist, lässt sie sich möglicherweise noch weiter verbessern. So musste der Studierende z.B. in der Studie von Clifton (2005) vor der Bestätigung jeweils zunächst entscheiden, ob er die Musterlösung verstanden hatte oder nicht.

### **Vor- und Nachteile der beiden Übungsmethoden**

Im Gegensatz zu praktikablen Tests mit Rückmeldungen sind Musterlösungen nicht an technische Geräte gebunden, lassen sich somit einfacher herstellen und vielfältiger, z.B. auch in einer Papierversion nutzen. Besonders nützlich dürften sich Musterlösungen für die Prüfungsvorbereitung eignen, da dann ein hohes Interesse und eine entsprechende Lernmotivation seitens der Studierenden meistens vorausgesetzt werden kann. Wegen der höheren Lerneffizienz können die Studierenden durch Musterlösungen mehr Lehrstoff bewältigen. Wenngleich Tests die Stärken, Schwächen und den Leistungsstand eines Studierenden objektiv erfassen und zuverlässige Rückmeldung dazu geben, kann daraus nicht so einfach gefolgert werden, Musterlösungen wären gänzlich ungeeignet, ähnliche Erkenntnisse beim Studierenden auszulösen.

### **Subjektive Einschätzungen beider Methoden**

Studierende halten Tests mit Feedback für lernwirksamer und insgesamt für die bessere Übungsmethode. Insofern dürfte ein Quizangebot im Vergleich zu Musterlösungen die Evaluation der Lehre aus Sicht der Studierenden qualitativ erhöhen. Aus der positiven Qualitätsbewertung der Quiz darf man aber nicht einfach ableiten, Studierende würden Quiz dem Studieren korrekt gelöster Aufgaben stets vorziehen. Nimmt man jeden Notendruck aus den Übungen heraus, so präferieren Studierende eindeutig Quiz mit Rückmeldungen. Dieser motivationale Vorteil lässt vermuten, Testaufgaben mit nützlichem Feedback würden auch ohne explizite Zwangsmaßnahmen eher zum Lernen genutzt werden. Die Quiz verlieren etwas an Qualitätseinschätzung und deutlich an Attraktivität, wenn sie benotet werden bzw. wenn unbenotete Übungsformen wie Musterlösungen als echte Alternativen zur Verfügung stehen. Trotzdem wird ein Quiz auch unter einer Benotungsbedingung immer noch als signifikant bessere pädagogische Übungsmethode eingeschätzt.

### **Nutzungsvorschlag für Quiz**

Angesichts der hohen Belastung im Studium erscheinen mir zusätzliche benotete Quiz recht problematisch, es sei denn, die Seminarbenotung basiere vornehmlich auf diesen Quiz, womit sie dann aber eher eine Test- als eine Übungsfunktion übernähmen. Die Lernwirksamkeit freiwilliger Quizangebote ist bis heute nicht gesichert, was aber weniger heißen soll, die freiwilligen Nutzer würden von solchen Tests nicht profitieren, sondern eher aussagt, im Mittelwert eines Seminars könnten so kaum Vorteile ausgemacht werden, unter anderem auch deshalb, weil zu wenige Studierende ein derartiges Angebot konsequent nutzen. Leider ergab sich in den hier untersuchten Seminaren keine

Möglichkeit, die Studierenden durch den Anreiz von Bonuspunkten zu motivieren, die im Gegensatz zu einer Pass/fail-Benotung keinerlei Bedrohung nach sich ziehen würden. Vielleicht sollte man unter solch restriktiven Bedingungen verpflichtende, unbenotete Tests einfordern und die unzumutbaren Leistungen einiger weniger Studierender dabei in Kauf nehmen. Im Übrigen kann man jede Testaufgabe so gestalten, dass der Lerner selbst entscheiden kann, ob er sie als Test oder Musterlösung auffassen will.

## Literatur

- Anderson, J. L., Weywadt, C., & McDaniel, M. A. (2008, October). Repeated quizzing facilitates learning of core content in an undergraduate neuroscience course. Poster presented at the Harvard Medical School Education Day, Boston, MA. [PDF] [http://psych.wustl.edu/memory/TELC/Anderson\\_et al\\_Poster\\_2008.pdf](http://psych.wustl.edu/memory/TELC/Anderson_et al_Poster_2008.pdf) [8.3.2010]
- Clifton, K.S (2005) The Testing effect: using retrieval practice in the classroom. Thesis submitted to Marshall University In partial fulfillment of the Requirements for the degree of Master of Arts Psychology. <http://www.marshall.edu/etd/masters/clifton-karen-2005-ma.pdf> [19.10.2005]
- Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14, 215-235.
- Hodge, G. (2009). Pedagogical Best Practices: The Power of Quizzing. The Redesign Alliance Third Annual Conference Orlando • March 24 [http://www.thencat.org/RedesignAlliance/2009%20Conference/Presentation%20Slides%2009/Hodges\\_UNMPowerofQuizzingPanel.pdf](http://www.thencat.org/RedesignAlliance/2009%20Conference/Presentation%20Slides%2009/Hodges_UNMPowerofQuizzingPanel.pdf) [29.12.2009]
- Jacobs, B. (1981). Angst in der Prüfung - Beiträge zu einer kognitiven Theorie der Angstentstehung in Prüfungssituationen. R. G. Fischer-Verlag. Frankfurt/Main  
Online Angsttest: URN: urn:nbn:de:bsz:291-psydok-26924;  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2010/2692/>
- Jacobs, B. (2003). Feedback mit oder ohne eigene Aufgabenbearbeitung? URL <http://www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/direktesfeedback.htm> [09.07.2010].
- Jacobs, B. (2007). Gezieltes Studieren oder Aufgaben mit Feedback als Übungsmethoden - Erfahrungsbericht eines missglückten Internetexperimentes. <http://www.phil.uni-sb.de/~jakobs/wwwartikel/testread/testread.html>
- Jacobs, B. (2008a). 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/>  
Langfassung: <http://www.phil.uni-sb.de/~jakobs/wwwartikel/testread/testread2.html>
- Jacobs, B. (2008b). Was wissen wir über die Lernwirksamkeit von Aufgabenstellungen und Feedback. In: Thonhauser, J. (Hrsg.) 2008: Aufgaben als Katalysatoren von Lernprozessen. Waxmann, Münster, 99-113.
- Jacobs, B. (2009a) Leistungssteigerung durch Notendruck ? - Die Wirkung der Benotung auf die Studierleistungen in einem Seminar. URN: urn:nbn:de:bsz:291-psydok-25299  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2009/2529/>

- Jacobs, B. (2009b). 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](http://www.phil.uni-sb.de/~jakobs/wwwartikel/extrinsische_motivierung/cash4quiz.html)
- Jacobs, B. (2009c). Vergleich von Bedingung A im aktuellen mit Bedingung B in einem früheren Jahrgang.  
[http://www.phil.uni-sb.de/~jakobs/seminar/vpl/experiment/exkurs/vorexperiment\\_variante.htm](http://www.phil.uni-sb.de/~jakobs/seminar/vpl/experiment/exkurs/vorexperiment_variante.htm)
- Jacobs, B. (2010a). Tests als Maßnahmen zur Förderung von Studierleistungen ?  
 URN: urn:nbn:de:bsz:291-psydok-26022  
 URL: <http://psydok.sulb.uni-saarland.de/volltexte/2010/2602/>  
[http://www.phil.uni-sb.de/~jakobs/wwwartikel/tests\\_als\\_studierhilfen/tests\\_als\\_studierhilfen.pdf](http://www.phil.uni-sb.de/~jakobs/wwwartikel/tests_als_studierhilfen/tests_als_studierhilfen.pdf)
- Jacobs, B. (2010b). Leistungssteigerung ohne Notendruck ? -Die Wirkung verpflichtender, unbenoteter Quiz auf die Studierleistung  
 URN: urn:nbn:de:bsz:291-psydok-26899  
 URL: <http://psydok.sulb.uni-saarland.de/volltexte/2010/2689/>
- Jacobs, B. (2010c). Testfragen selbst beantworten oder Musterlösungen studieren ?  
 URN: urn:nbn:de:bsz:291-psydok-26934  
 URL: <http://psydok.sulb.uni-saarland.de/volltexte/2010/2693/>
- Jacobs, B. (2010d). Bestanden/nicht bestanden im Vergleich zum traditionellen, mehrstufigen Benotungssystem.  
[http://www.phil.uni-sb.de/~jakobs/wwwartikel/extrinsische\\_motivierung/notendruck/pass\\_fail.html](http://www.phil.uni-sb.de/~jakobs/wwwartikel/extrinsische_motivierung/notendruck/pass_fail.html)
- Kluger, A. N. & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin*, 119 (2), 254-284
- LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, 67, 259–266.
- Larsen, D. P., Butler, A. C., Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Medical Education*, 43, 1174–1181.
- McDaniel, M. A., Anderson, J. L., Derbish, M.H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19, 494-513.
- Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249-255.