

Autor: [Bernhard Jacobs](#), Fachrichtung 5.1 Erziehungswissenschaft der Universität des Saarlandes
created: 15.1.2009

URL des Originals: <http://www.phil.uni-sb.de/~jakobs/wwwartikel/teststudy/teststudy4.html>

Die Wirkung der Übungsverteilung beim Studieren und Testen auf das Behalten der Bundesstaaten der USA

Abstract

Studierende sollten sich die Zuordnung der Namen der Bundesstaaten der USA zu ihren Territorien einprägen. Nach einer Lernaneignungsphase wurden sie nach Zufall auf 4 verschiedene Übungsbedingungen aufgeteilt, die sich zum einen in der Übungsverteilung *massiert oder verteilt* und zum andern durch die Übungsmethode *Einprägen der korrekten Antwort* oder *Testen mit Feedback* unterschieden. Der eine Woche nach den Übungen angesetzte Behaltenstest ergab keine Unterschiede zwischen den Übungsmethoden, aber einen Vorteil für das verteilte Lernen, der sich jedoch nur bei der Übungsmethode Testen mit Feedback deutlich herauskristallisierte. Einige interessante Prozessdaten während der Übung belegen, dass die Studierenden in der gleichen Übungszeit deutlich mehr Aufgaben durch Einprägen als durch Testen bearbeiteten und beim Testen die verfügbare Zeit für Aufgabenbearbeitung und Feedbacknutzung im Prinzip sinnvoll einsetzten. Alle Übungsvarianten wurden hinsichtlich der pädagogischen Qualität und Motivierung sehr ähnlich eingestuft.

Schlagworte:

Übungsmethode, Übungsverteilung, Übungswiederholung, massiertes und verteiltes Lernen, Testen mit Feedback, Einprägen; massed practice, distributed practice, spacing, massed and spaced repetitions, testing with feedback, restudy

Einleitung und Problemstellung

Eine bereits auf Ebbinghaus zurück gehende und offenbar empirisch gut gestützte pädagogische Empfehlung lautet, man solle das mehrfache Einüben eines Wissensgebietes zeitlich strecken und die Wiederholungen auf mehrere Termine verteilen, um möglichst langfristiges Behalten zu fördern. Eine ebenfalls schon recht lang bekannte, wenngleich zwischenzeitlich in Vergessenheit geratene psychologische Erkenntnis besagt, das Testen selbst stärke das Behalten, ein angemessenes Feedback unmittelbar nach der Testung verbessere den Lernerfolg darüber hinaus und ein so konzipierter Übungstest könne erneutem Studieren überlegen sein. Ziel der nachfolgenden Studie war es, beide Behauptungen zugleich sowie die Kombination beider Methoden am Beispiel des Erlernens der 50 Bundesstaaten der USA zu überprüfen

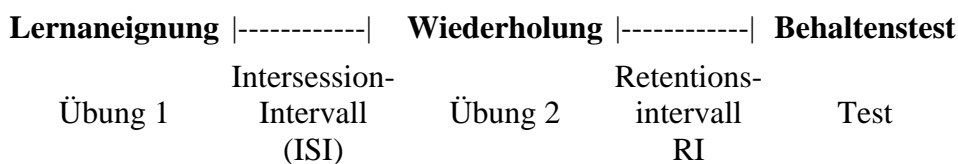
Verteiltes und massiertes Lernen als Übungsorganisation

Soll eine Liste von 20 Vokabeln gelernt werden, so besteht ein konventioneller Übungsdurchgang aus dem Einüben aller 20 Items. Wir schließen hier den pädagogisch eher uninteressanten Fall aus, die Wiederholung einer Vokabel bereits vor dem Einüben aller anderen Vokabeln zu beginnen, wie dies in vielen alltagspsychologischen Untersuchungen zum Spacing-Effekt (z.B. Pavlik & Anderson 2005) angenommen wurde. Denn es ist empirisch hinreichend geklärt, dass eine direkte permanente Wiederholung eines bestimmten Items (spacing lag=0) bzw. ein sehr kurzes inter item spacing lag langfristiges Behalten nicht fördert (Cepeda, Pashler, Vul, Wixted & Rohrer 2006). Unser Interesse gilt vielmehr ausschließlich der Wiederholung von Ü-

bungsdurchgängen. Ein Übungsdurchgang umfasst dabei die Bearbeitung aller einzulübenden Items. In dieser Untersuchung werden hierbei stets die 50 Bundesstaaten der USA in einer zufälligen Reihenfolge als Studier- oder Testvariante mit Feedback zum Einüben angeboten.

Werden 20 Vokabeln nun dreimal direkt hintereinander eingeprägt, also 3 Übungsdurchgänge verzögerungsfrei unmittelbar aneinander gereiht, so spreche ich im Gegensatz zu Cepeda et. al. 2006 von einer massierten Übung, die relativ gut unmittelbares Behalten fördert. Bei einer verteilten Übung im Sinne eines between session spacings oder inter session spacings liegt zwischen den Übungsdurchgängen ein deutliches Zeitintervall, das so genannte Inter-Study-Intervall bzw. Inter-Session-Intervall (ISI). Im Anschluss an die letzte Übung wird dann in einem bestimmten Zeitabstand der Lernerfolg überprüft. Hier interessiert ein etwas langfristigeres Behalten, genauer ausgedrückt, ein Retentionsintervall von einer Woche. Unter dem Retentionsintervall versteht man die Zeit zwischen dem letzten Übungsdurchgang und dem Behaltentest. Abbildung 1 zeigt den einfachsten Fall der zeitlichen Abfolge bei einer Wiederholungsübung mit Überprüfung

Abbildung 1. Schematische Darstellung einer Wiederholungsübung mit Behaltentest



Eine massierte Übungswiederholung entspricht einem ISI von 0. Während die verteilte Übung theoretisch ein ISI von größer 0 voraussetzt, sollte man in der pädagogischen Praxis ein deutliches Interstudierintervall ins Auge fassen. Mit wachsendem ISI steigt in der Regel die Behaltensstabilität. (siehe z.B.: Bahrick & Hall. (2005) bzw. [Die Wirksamkeit verteilter Übungen mit Tests und Feedback nach Bahrick und Mitarbeiter](#)) zumindest bis zu einem gewissen Optimum (siehe weiter unten). Daher werden deutlich bessere langfristige Behaltensleistungen bei verteilter Übung angenommen, was auch etliche Gesamtübersichten der empirischen Literatur erwarten lassen (z.B.: Dempster 1996, Donovan & Radosovich 1999, Cepeda et. al. 2006).

Der Behaltensvorteil verteilter Übung gilt gleichermaßen für reines Studieren wie für Testen und damit auch für Testen mit Feedback. Von besonderer Relevanz ist daher das Resümee von Dempster (1989) zu werten. "There is considerable evidence, gathered in a variety of settings and across many different types of materials and procedures, that spaced repetitions regardless of whether they are in the form of additional study opportunities or successful tests are a highly effective means of promoting learning."

Testen und Studieren als Übungsmethoden

Unzählige Studien ergaben deutliche Vorteile des Testens gegenüber gar nichts. Mit hin stellt bereit pures Testen (auch ohne Rückmeldung) unabhängig von seinem diagnostischen Wert eine effiziente Form von Übung dar, die in der Regel durch KCR-Feedback (= die korrekte Antwort) an Wirksamkeit gewinnt. (siehe Jacobs 2008b, S.

100-102). Von besonderem theoretischem wie praktischen Interesse ist die Beantwortung der Frage, ob Testen mit Feedback die Behaltensstabilität mehr fördert als erneutes Studieren auf der Basis der direkten Präsentation der korrekten Antwort. In den meisten Studien (z.B. LaPorte & Voss 1975, Cull 2000) konnte der Behaltensvorteil des Testens mit Feedback gegenüber dem Einprägen der korrekten Antwort nachgewiesen werden. In manchen Studien gelang der Nachweis eines Testvorteils nur bei bestimmten Testformen (z.B. nicht mit MC-, aber mit Short Answer Tests, etwa: Kang, McDermott & Roediger 2007) und gelegentlich auch gar nicht (z.B.: Jacobs 2008a, c), so dass hier insgesamt mit einem leichten Vorteil der Testens mit Feedback gerechnet wird.

Vermutete Wirkung von Übungsmethoden und Übungsorganisation auf das Behalten

Zu Beginn der Untersuchung lagen weder theoretische Erkenntnisse noch Erfahrungen hinsichtlich einer Behaltenswirkung für die Kombination von Übungsverteilung und Übungsmethode zugrunde. Es ist müßig darüber zu spekulieren, ob sich positive Effekte kumulieren oder neutralisieren bzw. ein positiver Faktor, z.B. Übungsverteilung die relative Schwäche des anderen Faktors (z.B. reines Einprägen) egalisiert. Intuitiv hat es den Anschein, eine massierte Übung in Form des Einprägens der korrekten Antwort hinterlasse relativ betrachtet ein besonders schlechtes langfristiges Behalten. Auch die Gewichtung der beiden Faktoren Übungsform und Übungsverteilung ist schwer einzuschätzen. Ich schätze die Effektstärke zugunsten des Testens mit Feedback im Vergleich zum Einprägen der korrekten Antwort als ziemlich klein bis höchstens mittel ein, habe aber zu wenig Erfahrung, um das Ausmaß der Übungsverteilung für die hier anstehende Versuchsanordnung quantitativ bewerten zu können. Ohne dezidierte theoretische Begründung und empirische Fundierung erwartete ich in Analogie zu Befunden von Cull 2000, dass die verteilte Übung durch Testen mit Feedback am besten, massierte Übung durch Einprägen am schlechtesten und die restlichen Methoden irgendwo dazwischen liegen.

Erst nach Durchführung der Studie wurde ich auf die Dissertation von Arne Weigold (2008) aufmerksam, welche den Einfluss der Übungsmethode Testen und Studieren in Kombination mit dem Spacings einer Übungswiederholung auf das Behalten von Vokabeln analysierte. Der Untersuchungsansatz unterscheidet sich jedoch im mehrfachen Weise bedeutsam von der hier dargestellten Studie, da das ISI zwischen einem und 13 Tagen variierte, bei der Lernaneignung ein Kriterium von 90% korrekter Antworten verlangt war, die Übungsmethode Testen bei der wiederholten Übung kein Feedback beinhaltete und das Retentionsintervall 13 Tage umfasste. Weigold (S.62, obere Graphik in der Figure 4) fand Behaltensvorteile des reinen Testens gegenüber dem Studieren für ein ISI von einem Tag und Nachteile bei einem ISI von 13 Tagen. Aus meiner Sicht ist das Ergebnis insofern verständlich, als die Erfolgswahrscheinlichkeit beim Testen nach einem Tag deutlich höher als nach 13 Tagen ausfällt und reines Testen lediglich das Behalten korrekt gelöster Items fördert. Ein erneuter Studierdurchgang nach 13 Tagen verhindert zwar den lernfördernden Abruf gekannter Items, erlaubt jedoch die verbesserte Kodierung vergessener Items, was offenbar mehr Lerneffektivität nach sich zieht. **Der Vergleich gezieltes Studieren gegen Testen mit Feedback scheint mir fairer und pädagogisch relevanter zu sein. Testen mit Feedback ist einfachem Testen ohne Rückmeldung klar überlegen und es macht pädagogisch keinen Sinn, eine schwächere Methode einzusetzen, wenn eine offensichtlich bessere verfügbar ist.**

In der Zwischenzeit untersuchten Carpenter, Pashler & Cepeda (2008) die Wirkung einer extremen Übungsverteilung (1 Tag vs. 4 Monate) beim Studieren und Testen (mit Feedback) historischer Fakten auf sehr langfristiges Behalten (9 Monate). Sie fanden Vorteile für Testen mit Feedback und verteiltes Lernen, aber keine Interaktion zwischen Übungsmethode und Übungsverteilung.

Experimentelle Bedingungen und Messvariablen

Die unabhängigen Variablen

Entsprechend dem Ziel, die Auswirkungen der Übungsvarianten und der Übungsverteilungen gleichzeitig zu untersuchen, wird das Experiment varianzanalytisch als zweifaktorielles Design betrachtet und umfasst als unabhängige Variablen die Faktoren:

Übungsverteilung mit den Stufen massiert (ISI=0) und verteilt (ISI=24 Stunden), Übungsmethode mit den Stufen gezieltes Studieren (study only) und Testen mit Feedback

Die genaue Ausgestaltung der Stufen der Übungsverteilung wird beim Versuchsablauf näher expliziert. Die Übungsmethoden lehnen sich an bereits erprobte Varianten von Jacobs 2006a an. Hierbei bestimmt der Lernende selbst sein eigenes Lerntempo. Tabelle 1 beschreibt beide Übungsmethoden und verweist auf entsprechende Links zu den Übungsprogrammen. Die lauffähigen Übungsprogramme beziehen sich auf die massierte Übungsbedingung und begrenzen die Übungszeit auf 14 Minuten. Die entsprechenden Übungsprogramme der verteilten Übungen sind identisch und unterscheiden sich lediglich durch eine auf 7 Minuten begrenzte Übungszeit.

Tabelle 1: Beschreibung der Übungsmethoden (hier für die massierten Übungen)	Userkennung: Alabama Passwort: Alabama
Gezieltes Studieren Study only (SO): Der Lerner sieht zunächst die Landkarte der USA ohne Staatsnamen. Durch Tastendruck wird ein Name im zutreffenden Staatsgebiet sichtbar. Bei erneutem Tastendruck verschwindet der alte Name und der nächste Staatsname erscheint im zutreffenden Bundesstaat. Die Reihenfolge der Staaten wird stets nach Zufall bestimmt. Nach sequenzieller Präsentation aller Bundesstaaten sieht der Lerner für eine Minute die Landkarte der USA mit den Namen aller Bundesstaaten. Anschließend werden die Staaten wieder einzeln bearbeitet. Der Ablauf wiederholt sich solange, bis die festgelegte Übungszeit erreicht ist.	<u>Präsentation der korrekten Lösung</u>
Testen mit Feedback Der Lerner sieht stets die Landkarte der USA ohne Staatsnamen. Er wird aufgefordert, das Territorium eines bestimmten Staates mit der Maus anzuklicken. Als Feedback folgt KOR (symbolisch richtig oder falsch), KCR (der geforderte Staatsname wird korrekt im Staatsgebiet platziert) sowie bei Fehler RCF (im falsch angeklickten Territorium wird der Name dieses Staates sichtbar). Die Reihenfolge der Staaten wird stets nach Zufall bestimmt. Nach sequenzieller Testung aller Bundesstaaten sieht der Lerner für 30 Sekunden die Namen aller falschen oder verwechselten Staaten in den zutreffenden Gebieten. Anschließend werden die Staaten wieder einzeln bearbeitet. Der Ablauf wiederholt sich solange, bis die festgelegte Übungszeit erreicht ist.	<u>Testen mit Feedback</u> <u>ck</u> <u>KOR,KCR,RCF</u>

Nach jedem Übungsdurchgang sahen die Probanden unter reinem Studieren die Landkarte der USA mit allen Staatsnamen, die Studierenden unter der Testbedingung die Landkarte mit den falschen bzw. verwechselten Staaten. Dadurch sollte die Übung etwas aufgelockert und zudem kurzfristig ein ergänzender Lernzugang ermöglicht werden, der für beide Übungsmethoden Vorteile erbringen sollte, ohne jedoch die jeweilige Methode im theoretischen Sinne abzuändern. Denn durch die Betrachtung aller Staaten der USA wird lediglich eine alternative Form gezielten Studierens angeregt und die Präsentation der falschen und verwechselten Staaten kann als besondere Form des

Feedbacks der Testgruppe gedeutet werden. Der überwältigende Anteil der Bearbeitungszeit (im Durchschnitt mindestens 80%) entfiel auf die Übungen innerhalb eines Übungsdurchgangs, das genaue Verhältnis beider Übungsteile variierte jedoch in Abhängigkeit von der Bearbeitungsgeschwindigkeit der einzelnen Items. Die Festlegung auf 7 Minuten Übungszeit bei der verteilten Übung leitet sich aus bisherigen Erfahrungen zu Testbearbeitungszeiten der Bundesstaaten ab und basiert auf der Intention, den meisten Übenden unter der Testbedingung wenigstens einen gesamten Testdurchgang zu ermöglichen.

Die Testungen entsprechen einem MC-Aufgabentyp mit 50 Antwortoptionen. Gegenüber früheren Versionen wurde bei der Itemsteuerung auf eine spezielle Flashcardmethode verzichtet. Das früher verwandte Flashcard mit Ausschluss korrekt gelöster Items im Übungsverlauf eines Übungsdurchgangs (siehe [Beispiel für das Erlernen einer Staatengruppe aus einer früheren Untersuchung](#)) bewirkt eine verstärkte Einübung fehlerhafter Items, weil falsch beantwortete Items stets an das Ende der Übungsliste angefügt und erst dann aus der Liste entfernt werden, nachdem sie korrekt beantwortet wurden. Damit werden die schwierigsten Items am stärksten geübt, was jedoch unter Zeitbeschränkung keine optimale Strategie darstellt. Bei beschränkter Übungszeit erweist es sich als sinnvoller, diejenigen Items verstärkt zu bearbeiten, die relativ leicht sind, die man aber noch nicht sicher beherrscht und bei denen durch Übung die Behaltenswahrscheinlichkeit am stärksten zunimmt. (z.B. Kornell & Bjork 2007). Da der Lerner hier die Bearbeitungszeit für jedes Item selbst bestimmen kann, bleibt es ihm überlassen, wie intensiv er sich mit den jeweils präsentierten Staaten beschäftigen will. Die spezielle Flashcardmethode erbrachte bisherigen Erfahrungen zufolge keine Behaltensvorteile gegenüber Testvarianten ohne spezielles Flashcard (Jacobs 2006a, 2007)

Abhängige Variablen

Als wichtigste abhängige Variable dient der Behaltenstest eine Woche nach der letzten Übung. Im Behaltenstest "**Bundesstaaten der USA zuordnen können**" werden alle Bundesstaaten der USA in zufälliger Reihenfolge getestet. Hierbei gibt der Computer jeweils den Namen eines Bundesstaates vor und fordert den Testanten auf das entsprechende Staatsgebiet mit der Maus anklicken. Die hohe Zuverlässigkeit des Testinstruments konnte in früheren Studien konsistent belegt werden. (Jacobs 2006a, 2007, 2008a). Darüber hinaus wurden etliche Prozessdaten erfasst, die Auskunft über das Nutzungsverhalten während der Übung geben sollten. Eine kurze Befragung am Ende der Übung bezieht sich auf die subjektive Bewertung der Übungsmethoden im Hinblick auf Lernwirksamkeit, Beanspruchung sowie Lernmotivation bzw. Interesse.

Versuchsdurchführung und Versuchsplan

Versuchspersonen

Als Versuchspersonen des ersten Experimentes dienten Studierende des Lehramts, die beim Verfasser dieser Arbeit ein Seminar im SS 2007 belegten. Den Studierenden wurde mitgeteilt, das Experiment sei Bestandteil des Seminars, diene als Veranschaulichung verschiedener Prinzipien (Diagnostik, Aufgabenformen, Lernmethoden) und seine Durchführung berechtige dazu, einer Seminarsitzung ohne Angabe von Gründen fernbleiben zu dürfen. Im zweiten Experiment nahmen Studierende des Lehramts teil, die im WS 2008/2009 an einem Seminar des Verfassers dieser Arbeit teilgenommen hatten. Ihnen wurde mitgeteilt, das Experiment sei Bestandteil des Seminars und die dafür investierte Zeit werde durch Ausfall sonstiger Seminarzeit kompensiert. Es soll-

ten nur solche Studierende teilnehmen, welche die Bundesstaaten der USA nicht oder kaum zuordnen könnten.

Versuchsablauf

Die Studierenden bearbeiteten den Versuch via Internet am eigenen Computer zu Hause. Sie erhielten vom Versuchsleiter eine Planvorlage, um sich Termine sowie den jeweils relevanten Link zum weiteren Vorgehen schriftlich zu notieren. Der Computer gab am Ende eines Versuchsteils jeweils Auskunft darüber, welches Programm der Studierende an welchem Tag zu absolvieren hatte. Es blieb den Studierenden überlassen, wann sie mit dem Experiment beginnen wollten, aber ein letztmöglicher Termin genannt.

Abbildung 2 stellt den Versuchsablauf rein schematisch dar, um einen direkten Vergleich der Bedingungen massierte und verteilte Übung zu ermöglichen.

Abbildung 2: Schema des Versuchsablaufs

Tag: Nr.	1				2	3		4	5	6	7	8	9	10
massive Übung	Lernaneignung Landkarte 10 min	Analogieaufgaben 6 min	Übung 14 min	Befragung 1 min								Behaltenstest		
verteilte Übung	Lernaneignung Landkarte 10 min	Analogieaufgaben 6 min			Übung 7 min	Übung 7 min	Befragung 1 min							Behaltenstest

Alle Probanden begannen am ihrem Tag 1 mit der Lernaneignungsphase. Als Lerngrundlage diente die Landkarte der USA mit den Namen der Bundesstaaten in den entsprechenden Regionen. Die Probanden sollten eine sinnvolle Strategie anwenden, um sich alle Bundesstaaten möglichst gut einzuprägen. Die Bearbeitungszeit wurde exakt auf 10 Minuten festgelegt, was den Studierenden auch mitgeteilt wurde. Diese Lernaneignungsvariante unterscheidet sich vom früheren Vorgehen (Jacobs 2006a, 2007, 2008a) durch die Präsentationsmethode sowie eine Zeitverkürzung um 4 Minuten.

Im Anschluss an die Lernaneignungsphase folgten 10 Analogieaufgaben, eine Minute Erholungspause sowie die Aufforderung zur Identifikation, was alles insgesamt ca. 6 Minuten Zeit beanspruchte. Sinn dieser Ablenkung war es, die massierte Übung zumindest etwas zu entzerren, um möglichen Frust und Langeweile, sowie die kognitiv eintönige Beanspruchung in noch vertretbaren Grenzen zu halten. Das Computerprogramm legte nun nach Zufall fest, welche der 4 möglichen experimentellen Bedingungen die TeilnehmerInnen zu absolvieren hatten. (siehe Versuchsplan)

Für die Probanden unter der Bedingung "verteilte Übung" war die Übung an diesem Tag beendet, während die TeilnehmerInnen unter der massierten Übung nun 14 Minuten ohne Unterbrechung entweder die reine Studiervariante oder die Testvariante mit Feedback bearbeiteten. In der abschließenden kurzen Befragung bewerteten die Studierenden ihre Treatmentvariante anhand einiger Adjektive (z.B. interessant, behal-

tenswirksam). Die verteilten Übungen entsprachen inhaltlich und von der Übungszeit exakt den massierten Übungen. Sie begannen lediglich einen Tag nach der Lernaneignung und verteilten sich auf 2 Übungszeitpunkte, die einen Abstand von jeweils einem Tag umfassten. Der Behaltenstest "Bundesstaaten der USA zuordnen können" fand für alle am Experiment beteiligten Personen eine Woche nach ihrem letzten Übungstermin statt. Tabelle 2 schematisiert den Versuchablauf etwas anders und vereinfacht ihn auf die wesentlichen Kernpunkte der Übungsverteilung.

Tabelle 2. Vereinfachte Darstellung der Übungsverteilung massiert vs. verteilt

	Tag 1	Tag 2	Tag 3	Tag 10
massiert			massierte Übung	Behaltenstest
verteilt	Übung 1	Übung 2	Übung 3	Behaltenstes

Bei dieser Deutung wird auch die für alle Treatmentvarianten identische Lernaneignung als Übung aufgefasst. Unter verteilter Übung finden dann 3 Übungssitzungen im Abstand von jeweils einem Tag statt, während die massierte Übung zu einem Übungstermin 24 Minuten konzentriertes Einprägen der Bundesstaaten der USA verlangt. Die rein experimentellen Übungsmethoden umfassen jeweils einen Anteil von 14/24 der gesamten Lernprozedur.

Bei der massierten Übung ist das Interstudierintervall (ISI) pragmatisch betrachtet 0, während es bei den verteilten Übungen im Durchschnitt einen Tag bzw. 24 Stunden umfasst. Das Retentionsintervall beträgt für jede Bedingung 7 Tage. Nach Rohrer & Pashler (2007) sowie Pashler, Rohrer, Cepeda & Carpenter 2007 hängt das Ausmaß des Behaltenseffektes des verteilten Übens von der Beziehung zwischen ISI und dem Retentionsintervall (RI) ab. Normalerweise steigt der Effekt des verteilten Lernens mit wachsendem RI an, erreicht dann ein Optimum und fällt anschließend wieder ab, wenngleich der Lernvorteil gegenüber massiertem Üben langfristig erhalten bleibt. Der höchste Behaltensvorteil ist bei einem ISI/RI Verhältnis zwischen .10 und .20 zu erwarten. "...final-test memory is optimized for an ISI that is about 10%-20% of the RI." (Pashler, Rohrer, Cepeda & Carpenter 2007, S. Fig. 2. S. 189). Die hier gewählten Bedingungen kommen sehr nahe an dieses Optimum heran, da das Verhältnis zwischen ISI und RI hier $1/7 = .14$ beträgt.

Versuchsplan

Wie beim Versuchsablauf beschrieben, wurden die Probanden nach Zufall auf die 4 Bedingungen aufteilt. Somit liegt ein klassischer experimenteller Randomisierungsplan ohne Vortest zugrunde, der varianzanalytisch als vollständig überkreuzter 2x2 faktorieller Plan gedeutet wird (siehe Tabelle 3). Wie Jacobs (2008a) durch die Erhebung eines Vortest feststellte, war das Wissen über die Bundesstaaten USA bei einer der hier untersuchten Probandenklientel ähnlichen Studentengruppe ziemlich gering, da dort nur ca. 12 % aller Staaten vor Beginn der Übung erkannt wurden. Die Erhebung eines Vortests ergab auch keine nennenswerte statistische Effizienzsteigerung durch die Analyse von Lernzuwächsen.

Tabelle 3: Formalisierung des Versuchsplans

R	X _{1,1}	O
R	X _{1,2}	O
R	X _{2,1}	O
R	X _{2,2}	O

X _{1,1}	=	massiert, studieren
X _{1,2}	=	massiert, testen mit Feedback
X _{2,1}	=	verteilt, studieren
X _{2,2}	=	verteilt, testen mit Feedback

Um die Effizienz der Randomisierung zu prüfen, wurden eine Varianzanalyse mit den 4 Bedingungen als Faktorstufen und den abhängigen Variablen erfragter Abiturnotendurchschnitt sowie Alter mit anschließenden multiplen Mittelwertsvergleichen durchgeführt. Als einziger signifikanter Effekt ergab sich hierbei in Experiment 1 ein besserer Abiturnotenmittelwert der Bedingung massiertes Testen (1.9) gegenüber der Bedingung verteiltes Studieren (2.5). In Experiment 2 konnten hingegen keine signifikanten Unterschiede der Bedingungen im Abiturnotendurchschnitt festgestellt werden (2.2-2.5). In beiden Experimenten zeigten sich keinerlei Unterschiede im Alter (Experiment 1: 22-24 Jahre; Experiment 2: 21-22 Jahre). Die Randomisierung ist demnach im Experiment 2 voll geglückt.

Ergebnisse

Bedenken hinsichtlich der Seriösität der Leistungsdaten bei den Experimenten.

Alle bisherigen Versuche von Jacobs (2006a, 2007, 2008a) zum Erlernen der Bundesstaaten der USA fanden unter beaufsichtigten Laborbedingungen in den CIP-Räumen der Universität des Saarlandes statt. Bevor die Studierenden dort ihren Computerplatz eingenommen hatten, war das Experiment ordnungsgemäß installiert worden, wodurch bei allen Probanden von der medialen Darbietung her vergleichbare Bedingungen vorgelegen hatten. Die hier beschriebenen Experimente absolvierten die Studierenden demgegenüber via Internet in ihrer natürlichen Lernumgebung zu Hause.

Einige Studierende berichteten nun im Seminar, ihr Computer hätte - entgegen der Ankündigung - keinen Vollbildschirm ohne Statusleiste erzeugt. Ohne Vollbildschirm läuft das Experiment zwar wie geplant ab, bei der Testung der Staaten kann man jedoch beim Überfahren der Maus auf der Landkarte die Namen der Bundesstaaten in der Statuszeile erkennen und so die korrekte Antwort ohne näheres Nachdenken einfach übernehmen. Ein Studierender gab zu, bei der späteren Behaltenstestung der Versuchung dieser "unbeabsichtigten Hilfe" nicht zuverlässig widerstanden zu haben.

Im Durchschnitt erzielten die Studierenden des Experimentes 1 im Behaltenstest "Bundesstaaten der USA" 70% korrekte Lösungen und es fanden sich keinerlei Unterschiede zwischen den Bedingungen. Unter Laborbedingungen lagen die entsprechenden Prozentsätze bisheriger Erfahrung nach im Bereich von 39 bis 52%. Da der hohe Behaltenserfolg im Experiment 1 über jeder mir realistisch erscheinenden Erwartung lag und meiner Meinung nach höchstwahrscheinlich zu einem schwer quantifizierbaren Ausmaß durch "unerlaubte Hilfestellungen" zustande gekommen sein könnte, habe ich mich dazu entschlossen, auf die Behaltensdaten des Experimentes 1 vollständig zu verzichten. Diese Entscheidung führte letztlich dazu, das Experiment unter verbesserten Bedingungen zu replizieren. Dennoch erscheinen mir einige Daten aus Experiment

1 verwertbar, über die deshalb auch berichtet wird, da die Problematik sich hauptsächlich erst im Behaltenstest manifestierte.

Das in Experiment 1 erlebte Manko des unbeabsichtigten Eintrags der korrekten Antwort in der Statuszeile geht auf einen Programmierfehler des Microsoft Internet Explorers 6 (MIE) zurück, der den Befehl "fullscreen=yes...bzw. status=no" **bei vielen, aber nicht zuverlässig bei allen Rechnern** (bzw. Betriebssystemen-, Browserversionen oder Sicherheitseinstellungen) umsetzte. Glücklicherweise fand ich dann im Internet eine zwar umständliche, aber alternative Programmiervariante über den Befehl "ausführen..." , die im MIE zuverlässig den Vollbildmodus ohne Statuszeile erzwingt. Auf dieser Basis wurde dann das Experiment 2 via Internet durchgeführt. Auch bei Experiment 2 ist die Durchführungsobjektivität der Behaltensmessung sicherlich nicht voll erfüllt, da Methoden bedingt keinerlei direkte Kontrolle ausgeübt wurde, wenngleich durch zusätzlich erhobene Daten hinreichende Evidenz erlangt werden konnte, einige hoch verdächtige Leistungsergebnisse auszusondern (z.B. fast perfekte Leistung bei überlangen Testzeiten). Hier bleibt nur zu hoffen, dass durch die gelungene Zufallszuweisung der Probanden die interne Validität der UV nicht allzu sehr ramponiert wurde. Diese Unwägbarkeiten gehen allerdings zu Lasten einer höheren Fehlerstreuung und erschweren so den Nachweis eines Effektes.

Leistungsergebnisse im Behaltenstest, eine Woche nach der Übung in Experiment 2

Alle nachfolgenden Leistungsanalysen beziehen sich auf die Ergebnisse des zweiten Experimentes. 3 Probanden wurden zunächst vollständig aus dem Datensatz entfernt, weil sie entgegen den Anweisungen die Übungsmethoden mindestens 2 mal durchgeführt hatten und dabei teilweise unterschiedliche Bedingungen absolvierten. Einige wenige Ergebnisse des Posttests waren wegen unzureichender Identifizierung der vorausgehenden Übungsmethode (z.B. falscher oder fehlender Identifizierungscode) nicht weiter verwertbar.

Bei der augenscheinlichen Analyse der verbliebenen, als gültig betrachteten 80 Posttestwerten traten riesige Personenunterschiede zu Tage, die allerdings im Tukey-Boxplot keine Ausreißer hinterließen. Trotz der immensen Spannweite von 16 bis 100 Prozent korrekter Lösungen verblieben somit alle Probanden für die weiteren statistischen Rechnungen im Datensatz des Posttests. Begünstigt durch die große Personenvarianz, erzielte der Posttest "Bundesstaaten der USA zuordnen können" (BZ) eine Reliabilität von Cronbachs $\alpha = .94$, die etwas höher ausfällt als die bisher ermittelten Konsistenzen. (Jacobs 2006a, 2007, 2008a)

Die zweifaktorielle Varianzanalyse mit den Faktoren Übungsmethode (Studieren, Testen) und Übungsverteilung (massiert, verteilt) ergab keinerlei statistische Effekte der experimentellen Bedingungen. Wie aus der Tabelle 4 hervorgeht, erzielten weder die Hauptfaktoren, noch die Interaktion nennenswerte signifikante Effekte. Einzig der hier irrelevante konstante Term belegt die starke Personenabhängigkeit der Messungen.

Tabelle 4: Ergebnis der Varianzanalyse für den Behaltenstest

	F (1,76)	p
Konstanter Term	533.4	.000
Übungsverteilung	0.59	.45
Übungsmethode	0.19	.66
Interaktion	2.23	.14

Tabelle 5 präsentiert die Ergebnisse für die vier Bedingungskombinationen und die zwei Faktoren.

Tabelle 5: Prozentsatz der korrekten Lösungen im Nachtest,
eine Woche nach der letzten Übung

[Mittelwerte und Streuungen]

	testen			studieren			Übungsverteilung		
	M	S	N	M	S	N	M	S	N
massiert	54.2	17	18	64.4	29	21	59.7	24	39
verteilt	66.1	20	20	60.1	24	21	63.3	22	41
Übungs- methode	60.5	20	38	62.5	27	42			

Verteiltes Üben schneidet insgesamt numerisch marginal, aber eben nicht durchgängig signifikant besser ab als massiertes Üben. Lediglich beim Testen erbrachte verteiltes Üben hypothesenkonform deutlich höhere Behaltenswerte als massiertes Üben. Der Vorteil des verteilten Testens gegenüber dem massierten Testen lässt sich sogar durch einen unabhängigen t-Test bestätigen ($t(36)=1.93$, $p=0.031$; $d=.64$). Dafür gilt der Übungsgewinn beim verteilten Lernen hier offensichtlich nicht für das wiederholte Studieren, ein eindeutig der Erwartung diskonformes Ergebnis. Der knappe numerische Vorteil des Studierens gegenüber dem Testen widerspricht deutlich der angenommen Überlegenheit des Testens gegenüber dem Studieren. Insbesondere beim unmittelbaren Einüben nach der Lernaneignung sollte das Testen mit Feedback mehr Verhaltensstabilität erbringen als wiederholtes Einprägen.

In ihrer Gesamtheit legen die Leistungsergebnisse die Schlussfolgerung nahe, die Übungsmethode spiele bei der wiederholten Einübung des Zuordnens der Bundesstaaten der USA keine überragende pädagogisch bedeutsame Rolle und die Übungsverteilung zeige nur beim Testen den erwarteten Vorteil des verteilten Lernens.

Zweifel an den Leistungsdaten etlicher Studierender und revidierte Ergebnisse

Wie bereits erwähnt, lagen die durchschnittlichen Behaltenswerte in Experiment 1 bei "unrealistischen 70% korrekter Lösungen". Im hier ausgewerteten Experiment 2 betrug die durchschnittliche Behaltensleistung ca. 60%. Diese liegt immer noch deutlich über den 39 bis 52% korrekter Lösungen, welche unter vollständig kontrollierten Laborbedingungen erzielt worden waren (Jacobs 2006a, 2007, 2008a). Ein Teil dieser erhöhten Erfolgsquote könnte auf die veränderte Lernaneignungsphase zurückgehen. In den hier berichteten Experimenten sahen die Probanden zu Beginn die Landkarte der USA mit den Namen der Bundesstaaten. Diese Präsentationsform kommt den Lern-

gewohnheiten der Studierenden eher entgegen als die früher verwandte Methode einer seriellen Darbietung eines Bundesstaatennamens in seinem Staatsgebiet. Jedoch erschien der Verdacht, einige Studierende hätten beim Behaltenstest möglicherweise unerlaubte Hilfen benutzt, eben auch nicht ganz von der Hand zu weisen.

Der Zusammenhang zwischen Testergebnis und Bearbeitungszeit ergab ein signifikantes r von .40, was besagt: Je länger man den Posttest bearbeitete, desto besser fiel das Ergebnis aus. Darauf hin wurden mittels Boxplot von Tukey die Bearbeitungszeiten des Posttests analysiert und so etliche Ausreißer entdeckt. Schließlich habe ich festgelegt, alle Probanden, die den Posttest länger als 10 Minuten - und damit mehr als 2 Standardabweichungen über dem Mittelwert - bearbeitet hatten, aus dem Datensatz auszuschließen (siehe hierzu auch die Bearbeitungszeiten [Mczeit] unter kontrollierten Bedingungen bei Jacobs (2006a). Im Durchschnitt erzielten die so Ausgeschlossenen 88% korrekter Lösungen (Spannweite 80-96). Nach dem Ausschluss dieser 8 Probanden sackte der Zusammenhang zwischen Bearbeitungszeit und Posttestleistung auf insignifikante $r = .13$ ab. Die hohen Erfolgsquoten der Ausgeschlossenen waren deshalb verdächtig, weil korrekte Antworten in der Regel relativ schnell gegeben werden (siehe weiter unten bei Prozessdaten.). Es ist ziemlich wahrscheinlich, dass etliche dieser Studierenden unerlaubte Hilfsmittel benutzten, obgleich die Anonymität der Untersuchung keinerlei extrinsische Konsequenzen erwarten ließ.

Tabelle 6 präsentiert die Ergebnisse für die vier Bedingungskombinationen und die zwei Faktoren nach Ausschluss von 8 verdächtigen Probanden.

Tabelle 6: Prozentsatz der korrekten Lösungen im Nachtest
(nach Ausschluss von 8 Probanden)

[Mittelwerte und Streuungen]

	testen			studieren			Übungsverteilung		
	M	S	N	M	S	N	M	S	N
massiert	51.0	16	16	55.6	27	16	53.3	22	32
verteilt	65.2	20	19	60.1	24	21	62.8	22	40
Übungs- methode	58.7	20	35	58.4	26	37			

Auch diese Analyse bestätigt die vergleichbare Behaltenswirksamkeit von wiederholtem Studieren und Testen mit Feedback, spricht aber deutlicher für den Vorteil des verteilten gegenüber dem massierten Üben. Denn die Übungsverteilung zu Gunsten von verteiltem Lernen konnte jetzt sowohl insgesamt ($t(70)=1.78$, $p=.04$), wie auch nur für Bedingung Testen ($t(30)=2.27$, $p=0.015$) signifikant bestätigt werden.

Prozessdaten

Während der Übung wurden einige Daten erhoben, die nähere Auskunft über den Übungsablauf liefern und so die Nutzung einzelner Bedingungen genauer beschreiben sollten.

Die Übungszeit für alle Bedingungen war absolut vergleichbar. Lediglich die Übungsmethoden und die Übungsverteilung sollten variieren. Theoretische Überlegungen so-

wie die Ergebnisse früherer Untersuchungen ließen deutlich mehr Übungstrials bei der Methode "reines Studieren" (Study only) erwarten. Die Präsentation der korrekten Lösung beim Studieren korrekter Antworten erfordert kein vorheriges Erinnern und umfasst nur einen Teil von Testen mit Feedback. Wie aus Tabelle 7 hervorgeht, bearbeiteten die Studierenden unter Study only fast doppelt so viele Items wie unter der Methode Testen mit Feedback.

Tabelle 7: Anzahl der bearbeiteten Items in der jeweiligen Übung
[Mittelwerte und Streuungen]

		Experimentelle Bedingung		
		massiert	verteilt	
			Ü1	Ü2
Exp 1	Studieren korrekter Antworten (N 15-18)	192,5 82,0	98,8 26,7	103,8 50,0
Exp 2	Studieren korrekter Antworten (N 23-28)	181,7 77,7	103,5 51,7	125,6 65,5
Exp 1	Testen mit Feedback (N 15-18)	112,7 40,0	54,7 16,8	64,9 20,4
Exp 2	Testen mit Feedback (N 23-28)	111,9 30,8	56,3 23,2	64,0 26,2

Die Unterschiede der Itembearbeitungen zwischen Studieren und Testen mit Feedback sind so deutlich, dass sich eine Signifikanztestung erübrigt. Die Effektstärken d liegen bei allen möglichen Vergleichen jeweils deutlich über 1. Aus der Gruppe Studieren korrekter Antworten musste ein Ausreißer mit 500 Itembearbeitungen ausgeschlossen werden. Dieser klickte sich offenbar so schnell wie möglich durch alle Items, um dann mit Hilfe der Landkarte zu lernen, die nach jedem Übungsdurchgang eine Minute zur Verfügung stand.

Wie aus der Tabelle 7 ersichtlich, bearbeiteten die Studierenden unter der Bedingung verteilter Übung insgesamt fast annähernd gleich viele Items wie unter der Bedingung massierter Übung. (Experiment 1: Studieren: 203 vs. 193; Testen 119 vs. 113; Experiment 2. Studieren 182 vs. 229; Testen 112 vs. 120). Hinsichtlich der Übungsverteilung liegen somit theoretisch sehr günstige Bedingungen vor, da für jede Verteilungsstufe Arbeitszeiten und Übungstrials hoch vergleichbar ausfallen und nur der Abstand zwischen den Übungszeitpunkten variiert.

Behalten im Verlauf der Übungen

Ziel der Lernaneignungsphase war es, durch eine geeignete Präsentation der Information ein gewisses Lernfundament zu legen und Erfolgswahrscheinlichkeiten im Bereich von ca. 50% zu erzeugen. Übungen haben den Sinn, nach einer Lernaneignung das Lernen zu stabilisieren. Pädagogisch ernst zu nehmende Übungsmethoden dürfen nicht nur von rein theoretischem Interesse, sondern müssen praktisch an sich förderlich sein. Etliche mir bekannte Test- und Feedbackexperimente begannen z.B. ohne vorherige Instruktionsphase direkt mit einer Testung. So kann man zwar hohe Effekte produzieren, weil man quasi von einem Nullwissen ausgeht, wird aber kaum einen Pädagogen überzeugen, der in der Regel erstmal informiert, bevor er an Testung denkt. Zugleich muss ein Versuch jedoch hinreichenden Spielraum für nachfolgende Lernerfolgsunterschiede ermöglichen.

Insbesondere die verteilten Übungen waren zeitlich so festgelegt worden, dass alle Staaten der USA von möglichst vielen Studierenden mindestens einmal bearbeitet bzw. getestet werden konnten, was auch überwiegend, aber nicht immer gelang. Nachfolgende Tabelle 8 stellt den Prozentsatz korrekter Lösungen für die Bedingung Testen mit Feedback im jeweils ersten Übungsdurchgang einer Übung dar. Die Prozentsätze in den ersten beiden Spalten (T1 = Test 1) messen in gewisser Weise das Ergebnis der ersten Lernaneignungsphase, überschätzten es aber etwas, weil während der Übung noch hinzugelernt wird. Zum einen grenzt das KCR-Feedback gegenüber reiner Testung mögliche Falschantworten im Verlauf der Testung ein und das Response Contingent Feedback (RCF = die Mitteilung eines verwechselten Staates) ist teilweise als erneute Studiermöglichkeit zu werten. T2 gibt das Ergebnis des ersten Testdurchgangs der zweiten (bzw. letzten) Übung beim verteilten Üben wieder.

Tabelle 8: Prozentsatz korrekter Lösungen im jeweils ersten Übungsdurchgang einer Übung [**Mittelwerte** und *Streuungen*]

	Experimentelle Bedingung		
	Testen mit Feedback		
	massiert	verteilt	
	T1	T1	T2
Experiment 1	59,3	56,0	68,0
(N 15-16)	20,3	16,0	14,7
Experiment 2	63,1	49,7	60,1
(N 23-28)	13,9	21,7	24,2

Die bisherigen Ergebnisse (Jacobs 2006a, 2007) von jeweils 50% liegen deutlich niedriger als der direkt vergleichbare Prozentsatz von 59 und 63%, was möglicherweise an der veränderten Lernaneignungsphase liegt. Denn hier war die Landkarte der USA mit allen Namen der Bundesstaaten 10 Minuten zum Einprägen vorgegeben worden, was den Studiengewohnheiten vermutlich eher entspricht als die früher verwendete sequenzielle Präsentation von Staatsnamen in einem Staatsgebiet. Die Lernbedingungen zu Hause sind als angenehmer und weniger störanfällig einzuschätzen als die in einem großen Computerraum und erlauben möglicherweise ein besseres Einüben. Zudem wurde die Testung nicht unter Laborbedingungen kontrolliert und dann bleibt nicht ausgeschlossen, dass einige Studierende in ihren Notizen oder sonstigen Quellen nachgeschaut haben.

Die erste verteilte Übung durch Testen mit Feedback (T1 unter verteilt) fand einen Tag nach der Lernaneignung statt. Ich hatte etwas mehr Vergessen und somit einen geringeren Prozentsatz als ca. 50% erwartet. Der höhere Lernerfolg beim Test der zweiten Übung (T2 unter verteilt) von 68% bzw. 60 % wird durch den Lernzuwachs der ersten Übung verständlich. Der Lernzuwachs von T1 nach T2 wurde in beiden Experimenten jeweils durch einen t-Test für abhängige Stichproben eindrucksvoll signifikant bestätigt ($t(15)=3.25$, $p<.01$; $t(22)=3.86$, $p<.01$)

Die Angemessenheit des Response Contingent Feedback bei den Testgruppen

Im seinem Beitrag [Lohnt sich Antwort abhängiges Feedback ?](#) kommt Jacobs (2004) zu dem Ergebnis, spezielle Rückmeldungen zu bestimmten Fehlern seien dann durchaus zielführend, wenn die falsche Alternative ebenfalls zum Lehrziel gehört. Das ist hier der Fall, weil jeder falsch angeklickte Staat darauf verweist, der Lerner kenne we-

der den geforderten, noch den verwechselten Staat in hinreichendem Maße. Das RCF-Feedback kontrastiert den verwechselten Staat mit dem durch KCR-präsentierten verlangten Staat und fordert so zu einer vergleichenden Betrachtung beider Fehlerquellen, um diese in Zukunft zu vermeiden.

Tabelle 9: Prozentsatz der verwechselten Items
[Mittelwerte und Streuungen]

	Experimentelle Bedingung		
	massiert	verteilt	
		T1	T2
EX1	27,1	33,8	23,5
	15,7	13,3	11,1
EX2	28,5	20,8	16,4
	14,2	16,8	13,2

Aus Tabelle 9 erkennt man, dass ca. 1/5 bis 1/3 aller Antworten Verwechslungen darstellen, die insgesamt ein hinreichendes Verbesserungspotenzial beinhalten und so den Einsatz von RCF zumindest theoretisch als sinnvoll einschätzen lassen. Die Studierenden riskierten offensichtlich gerne einen Fehler, weil sie nur selten die Option "weiß nicht" anklickten. Die Testanweisung förderte jedoch diese Testhaltung. Um einen umfassenden Retrieval anzuregen und frühzeitiges Aufgeben nach Möglichkeit zu verhindern, erhielten die Probanden vor der Übung den Rat, möglichst intensiv nachzudenken und nur bei absoluter Unwissenheit auf "weiß nicht" zu klicken. Möglicherweise hat der Lerner häufiger die falsche Alternative nicht aus Überzeugung, sondern aus taktischen Erwägungen mehr oder weniger beliebig angeklickt, um möglichst zügig den korrekten Staat einsehen zu können.

Itemspezifische Test - und Feedbackbearbeitungszeiten

Kann ein Student einen bestimmten Staat relativ mühelos und sicher lokalisieren, dann trifft er vermutlich eine schnelle Entscheidung, während schwierige Staaten mehr Denkzeit erfordern. Im Folgenden werden korrekt gelöste Aufgaben eines Probanden als leicht, falsch beantwortete Aufgaben hingegen als schwierige Aufgaben aufgefasst. Für jeden Probanden wurde die durchschnittliche Bearbeitungszeit für seine korrekt gelösten wie für seine falschen Antworten erfasst und schnellere Antworten bei korrekten Lösungen erwartet. Unmittelbar nach der Itembeantwortung erhielt der Proband Feedback in Form von KOR, KCR und RCF. Es blieb ihm überlassen, wie lange er das Feedback einsehen wollte. Die Zeitspanne von der Bestätigung der Antwort bis zur aktiven Aufforderung der nächsten Aufgabe wurde als Feedbacknutzungszeit interpretiert. Sie wurde getrennt für korrekte und falsche Antworten ermittelt und die jeweiligen Durchschnittswerte eines Probanden verwendet. Bei korrekten Lösungen genügt eine kurze Bestätigung, für falsche Antworten ergibt sich die Notwendigkeit einer verbesserten Enkodierung, was einen deutlich längeren Zeitbedarf erfordert. Wie aus Tabelle 10 hervorgeht, bestätigen alle verfügbaren Daten die Vermutungen eindeutig, in hohem Ausmaß (d in der Regel mindestens 1) und konsistent zu allen erhobenen Übungszeitpunkten. Für alle möglichen Vergleiche gilt: Die Dauer von Itembeantwortung und Feedbacknutzungszeit fällt bei falschen Antworten stets hochsignifikant höher aus als für korrekte Antworten.

Tabelle 10: Durchschnittliche Bearbeitungs- und Feedbacknutzungszeiten pro Item [**Mittelwerte** und *Streuungen*]

		Experiment 1			Experiment 2		
		massiert	verteilt		massiert	verteilt	
		Ü	Ü1	Ü2	Ü	Ü1	Ü2
Itembeantwortungszeit							
	korrekte	5,4	5,2	4,2	4,9	5,0	4,6
	Antwort	3,3	2,7	2,2	2,7	2,7	2,7
	falsche	8,3	7,8	8,0	8,3	9,2	7,8
	Antwort	4,7	3,3	3,8	4,5	5,3	4,6
Feedbacknutzungszeit							
	korrekte	1,8	1,9	1,8	1,9	1,7	1,5
	Antwort	0,8	0,6	0,8	0,4	0,6	0,4
	falsche	3,4	3,4	3,4	3,3	3,1	3,0
	Antwort	1,1	1,3	2,5	0,7	1,4	1,1

Experiment 1: N pro Bedingung 15-16; Durchschnittliche Anzahl der Aufgaben massiert 113; verteilt ca. 60)

Experiment 2: N pro Bedingung 23-30; Durchschnittliche Anzahl der Aufgaben massiert 112; verteilt ca. 60)

Zugleich erkennt man an Tabelle 10 einen deutlich höheren Zeitbedarf für die Testung im Vergleich zur Feedbacknutzung. Die Ergebnisse stützten insgesamt ein vernünftiges Übungsverhalten durch ein den jeweiligen Anforderungen angemessenes Zeitmanagement. In etlichen Laborexperimenten zum Paarassoziationslernen wurden aus Gründen der experimentellen Vergleichbarkeit Beantwortungs- und Feedbackzeiten starr festgelegt (z.B. Cull 2000). Für das Lesen der Fragestellung und die Beantwortung lagen hierbei die zugestanden Zeiten deutlich höher als für das Feedback, was die Befunde in Tabelle 10 durchaus rechtfertigen. Die hier ermittelten Daten verdeutlichen jedoch zugleich, dass bestimmte Items mehr Präsentation und Feedbackzeit erfordern als andere Items. Außerdem signalisieren die Streuungen offenbar auch einen ziemlich unterschiedlichen individuellen Bearbeitungsbedarf. Eine für alle Personen und Items geltende konstante Zeitvorgabe, wie in der Experimentalpsychologie häufig festgelegt, wird den unterschiedlichen Zeitbedürfnissen der Studierenden offensichtlich nicht gerecht. Im Übrigen kann ich mir in der pädagogischen Praxis kaum vorstellen, Lerner würden sich durch starre Zeitvorgaben der einzuübenden Items widerstandslos unter Stress setzen lassen.

Subjektive Einschätzungen zur Übungsmethode

Nach dem Ende der jeweils letzten Übung sollten die Studierenden ihre Übungsmethode anhand einiger Items auf einer 7 stufigen Likertskala einschätzen. Die Skala war für jede Kategorie verbal verankert und erstreckte sich von "1 = stimmt überhaupt nicht" bis "7 = stimmt ganz genau". Bei den massierten Übungen bezog sich diese Einschätzung auf die letzten 14 Minuten, um eine Konfundierung mit der Lernaneignungsphase zu vermeiden. Der Einschätzung der Übungsmethode lag die Anweisung zugrunde. "Ich bewerte die letzten 14 Minuten des Übungsprogramms als ..." Unter verteilter Übungsmethode wurde auf diesen Zusatz verzichtet, da sich die Befragung auf den letzten Übungstermin bezog, der ausschließlich die verteilte Übung beinhaltete. 2 Probanden wurden wegen extremst möglicher negativer Einschätzung bei allen Items aus dem Datensatz von Experiment 1 entfernt.

Die meisten Items (siehe Tabelle 11) wurden zu nachfolgenden Fragebogentestwerten zusammengefasst, wobei die Testwerte selbst den Mittelwert aus allen verwandten Items repräsentieren. Der erste Alphakoeffizient bezieht sich auf die Reliabilitätsermittlung für das Experiment 1 (N=61), der zweite Koeffizient auf eine gesonderte Erfassung für die Daten von Experiment 2 (N=93)

Tabelle 11: Subjektive Einschätzungen und zugrunde liegende Items

Qualität der Unterrichtsmethode ($\alpha = .84$ sowie $\alpha = .85$):

gute Übungsmöglichkeit,
geeignete Übungsmethode,
pädagogisch empfehlenswert,
behaltenswirksam.

Lernmotivation/Interesse ($\alpha = .86$ sowie $\alpha = .83$)

interessant,
motivierend,
anregend,
herausfordernd,
langweilig(-).

Negative Einschätzung: ($\alpha = .84$ sowie $\alpha = .85$)

stupider Drill,
einfallsloses Pauken.

Die Reliabilitäten der Einschätzskalen fallen für beide Experimente sehr konsistent und hoch zufriedenstellend aus. Sie entsprechen zudem von der Höhe den bei Jacobs (2008) ermittelten Konsistenzen. Wenngleich sich die Skalen inhaltlich theoretisch voneinander abgrenzen lassen, so sind sie statistisch nicht unabhängig voneinander, da sie im Durchschnitt dem Betrage nach .71 bzw. .67 miteinander korrelieren und faktorenanalytisch einen Faktor bilden.

Hypothesen zur subjektiven Einschätzung und Ergebnisse

Für alle Skalen war eine positivere Einschätzung der Testmethode gegenüber der reinen Studierform erwartet worden, zumal entsprechende Präferenzen für die Testmethoden bereits bei Jacobs (2006a und 2008a) nachgewiesen werden konnten. Hinsichtlich der Übungsverteilung lagen keine klaren Erwartungen zugrunde. Wie die Ergebnisse in Tabelle 12 aufzeigen, liegen alle Mittelwertsunterschiede in Experiment 1 zwar in der erwarteten Richtung, fallen insgesamt aber zu schwach aus, um die Signifikanzhürde zu nehmen. Die Ergebnisse des Experimentes 2 können die Erwartungen ebenfalls nicht stützen. In der Gesamtbewertung komme ich zu dem Schluss, die Studierenden schätzten die beiden Übungsmethoden Studieren und Testen subjektiv annähernd vergleichbar im Hinblick auf Interesse oder Lernförderung ein.

Tabelle 12: Ergebnisse zur subjektiven Einschätzung der Übungsmethoden
[**Mittelwerte** und *Streuungen*]

Exp		Testen		Studieren			
1	Qualität der Übung	5.2	1.2	4.9	1.0	t(59)= 0,9	p=.36
2	Qualität der Übung	4.6	1.0	4.9	1.1	t(91)= -1,7	p=.09
1	Lernmotivation/Interesse	4.8	1.2	4.4	1.1	t(58)= 1.5	p=.13
2	Lernmotivation/Interesse	4.6	1.2	4.2	1.0	t(90)= 1.6	p=.11
1	negative Einschätzung	2.8	1.4	3.2	1.6	t(59)= -1.1	p=.29
2	negative Einschätzung	3.6	1.5	3.4	1.4	t(93)= 0.8	p=.45

Die Mittelwerte der Einschätzvariablen für massierte und verteilte Übungen (nicht in der Tabelle 12 enthalten) fallen ebenfalls sehr vergleichbar aus und lassen in beiden Experimenten keinerlei Trend für eine Form der Übungsverteilung erkennen.

Die nur durch eine kurze Auflockerung unterteilte massierte Übung erforderte 24 Minuten konzentriertes Einprägen, während die verteilten Übungen um ein Drittel kürzere Lernepisoden und diese verteilt auf 3 Tage umfassten. Objektiv erscheint die massierte Übung somit eindeutig anstrengender, was sich auch in der Einschätzung des Items "anstrengend" niederschlagen sollte. Eine zweifaktorielle Varianzanalyse mit den Faktoren Übungsaufteilung und Übungsform ergab in Experiment 1 für die Einschätzung "anstrengend" lediglich einen signifikanten Effekt und zwar für den Hauptfaktor Übungsaufteilung. Als Testgröße wird hier der Einfachheit halber der t-Test verwendet.

Tabelle 13: Mittelwerte und *Streuung* zum Item "Ich bewerte die Übungsmethode als anstrengend" (1= stimmt überhaupt nicht 7 = stimmt ganz genau) für Experiment 1 und 2

Experiment	massiert		verteilt				
1	5.1	1.2	3.4	1.6	t(59)= 4,6	p<.01	d=1.18
2	4.7	1.6	3.8	1.4	t(90)= 2,9	p<.01	d=0.60

Massierte Übungen wurden in Experiment 1 unabhängig von der Übungsmethode somit deutlich anstrengender eingeschätzt als verteilte Übungen. Ebenso schätzen die Studierenden massierte Übungen zumindest in Experiment 1 ermüdender ein als verteilte Übungen, wenngleich der Unterschied weniger deutlich, aber immer noch knapp signifikant (t(59)=2, p=.05) ausfällt. Bei Jacobs 2008a schätzen die Studierenden die ausschließliche Studiermethode anstrengender ein als eine Testmethode, was hier nicht bestätigt werden konnte. Allerdings diente bei Jacobs 2008a als Study only Variante die Landkarte und als Testvariante eine Form des selbst gesteuerten Testens. Die Situation in Experiment 2 sieht etwas komplizierter aus, wie die Mittelwerte der Bedingungskombinationen in Tabelle 14 verdeutlichen.

Tabelle 14: Zellenmittelwerte zum Item "Ich bewerte die Übungsmethode als anstrengend" (1= stimmt überhaupt nicht 7 = stimmt ganz genau) für Experiment 2

	Testen	Studieren
massiert	5.4	4.1
verteilt	4.1	3.6

Varianzanalytisch erzielten sowohl die Übungsmethode im Sinne "Testen ist anstrengender als Studieren" ($F(1,88)=8,6$ $p < .01$) wie die Übungsverteilung im Sinne "massiert ist anstrengender als verteilt" ($F(1,88)=8,9$ $p < .01$) signifikante Haupteffekte, während die Interaktion nicht signifikant ausfiel.

Zusammenfassung und Diskussion

Studieren und Testen mit Feedback sind vergleichbar gute Übungsmethoden

Die Daten von Experiment 2 sprechen in ihrer Gesamtheit ziemlich deutlich für die Gleichwertigkeit der Übungswiederholungsmethoden erneutes Studieren und Testen mit Feedback, was der generellen Forschungstendenz des Testvorteils zunächst einmal widerspricht. Zwar fand auch Jacobs beim Erlernen der Bundesstaaten gelegentlich einige Behaltensvorteile für Testen mit Feedback, häufig aber auch sehr ähnliche Lernerfolge im Vergleich zu reinem Studieren. In der Forschungsliteratur hatte man häufiger dann keine Unterschiede zwischen Testen mit Feedback und erneutem Studieren gefunden, wenn als Testmethode das MC-Format zum Einsatz kam. Auch hier lag als Testform MC zugrunde, allerdings in der ungewöhnlichen Aufgabenstellung mit jeweils 49 Distraktoren. Durch Zufall wird der Unwissende daher wohl kaum den korrekten Staat anklicken können. Auch ein triviales Wiedererkennen im Sinne von Recognition erscheint mir bei 50 Optionen unwahrscheinlich, zumal Jacobs 2006b einen sehr hohen Zusammenhang ($r = .88$ ($N=54$)) zwischen dem hier verwandten MC- und einem Short-Answer Format bei der Behaltenstestung der Bundesstaaten der USA gefunden hatte.

Die von Jacobs durchgeführten Experimente unterscheiden sich von den sonstigen Forschungsansätzen durch die selbstbestimmbare Bearbeitungszeit eines Items bei festgelegter Gesamtübungszeit. Dieser Ansatz ermöglicht eine optimale adaptive Übungsnutzung für jeden Lerner bei jeder Methode und kommt dem natürlichen Lernen sehr nahe. Schnelle Lerner können in derselben Zeit mehr lernen als langsame Lerner und das Studieren lässt sich in einer gegebenen Zeit eben häufiger anwenden als ein Testen mit Feedback. Bei etlichen Grundlagenexperimenten gehen diese Vorteile verloren, weil itemspezifisch für jedes Individuum und jede Methode eine konstante Zeit festgelegt wird (z.B.: 10 Sekunden für jedes Item beim Einprägen bzw. 6 Sekunden Testen, 4 Sekunden Feedback). So aber werden dann etwa gute Lerner gelangweilt, schwächere Lerner überfordert, leichten Items zu viel Zeit, schweren Items zu wenig Zeit gewährt und eine Übungsmethode im Durchschnitt länger als notwendig angewendet. Verzichtet man auf eine Festlegung der Gesamtübungszeit, hält die Anzahl der bearbeiteten Items konstant und gewährt eine itemspezifische Bearbeitungsfreiheit, dann müsste die Gesamtübungszeit für Testen mit Feedback höher als die für das einfache Studieren ausfallen und ein potenzieller Lernvorteil des Testens ist dann schwer interpretierbar. In einem pädagogisch anspruchsvollen Übungsexperiment im realen Universitätsbetrieb von Jacobs 2008c, welches Testen mit Feedback gegen reines Stu-

dieren überprüfte und hierbei die Bearbeitungszeit vollständig den Studenten überließ, erzielten beide Methoden vergleichbarer Übungseffekte, das reine Studieren jedoch in kürzerer Übungszeit. Dieses Ergebnis war interpretierbar und zwar im Sinne höherer Lerneffizienz für reines Studieren im Vergleich zu Testen mit Feedback.

Die entsprechenden Unterschiede im experimentellen Aufbau sind bei der Interpretation der Ergebnisse zu berücksichtigen und können durchaus zu verschiedenen Resultaten führen. Wenn wie hier Testen mit Feedback "erwartungswidrig" den gleichen Lernerfolg erbrachte wie Einprägen der korrekten Antwort, dann könnte ich das anders auch wie folgt ausdrücken: "Um den gleichen Effekt beim Einprägen wie beim Testen mit Feedback zu erzielen, mussten die Items ca. doppelt so oft eingeprägt werden. Einprägen ist hier deshalb gleich gut wie Testen mit Feedback, weil das doppelte Einprägen eben in derselben Zeit möglich war."

Ich komme auf der Basis der bisherigen Befunde beim Einprägen der Bundesstaaten der USA zu der Schlussfolgerung, die möglichen Unterschiede der beiden Übungsmethoden seien unter fairen Bedingungen letztlich zu gering und entbehrten hinreichender praktisch pädagogischer Bedeutsamkeit, um der Präferenz einer bestimmten Variante die klare Priorität verleihen zu können.

Verteiltes Lernen teilweise besser als massiertes Lernen

Das massierte Lernen nahm hier Ausmaße an, welche die Geduld der Studierenden schon etwas beanspruchte. Somit verwundert es nicht, dass die massierte Übung anstrengender erlebt wurde als das verteilte Üben. Die höhere Anstrengung des massierten Lernens geht allerdings fast ausschließlich auf das massierte Testen zurück, welches tendenziell auch die schwächsten Behaltenswerte nach sich zog. Inter-Study-Intervall (ISI) und Retentionsintervall waren so gewählt worden, damit sich ein größtmöglicher Vorteil für das verteilte Lernen manifestieren sollte. Aber lediglich bei der Übungsmethode Testen mit Feedback ließ sich der erwartete Behaltensvorteil des verteilten Lernens gegenüber dem massierten Lernen sowohl für alle Daten, wie für die bereinigten Daten statistisch eindeutig und in akzeptabler Höhe sichern. Bei der Übungsmethode wiederholtes Studieren deutete sich die Überlegenheit des verteilten Lernens lediglich bei den revidierten Daten etwas an, war aber insgesamt zu schwach, um die Signifikanzhürde zu nehmen.

Keine klaren subjektiven Präferenzen für eine bestimmte Übungsmethode.

Es war insgesamt nicht gelungen, klare Unterschiede in der subjektiven Bewertung der einzelnen Übungsmethoden aufzuspüren, wenngleich Präferenzen für die aufwändige Testung mit etlichen Rückmeldungen erwartet worden waren. Die Studierenden schätzten allerdings lediglich die ihnen zugewiesene Übungsmethode ein. Jedoch konnte Jacobs 2008a mit derselben Methode deutliche Unterschiede in der Qualitätseinschätzung zwischen einer speziellen Methode selbst gesteuerten Testens und dem Lernen mit der Landkarte feststellen. Subjektiv eingeschätzte Methodeneffekte sind vermutlich eher zu erwarten, wenn die Probanden über hinreichende Erfahrung mit allen Übungsvarianten verfügen würden. Bei Jacobs (2006a) bearbeiteten die Probanden mehrere Methoden in einem Wiederholungsdesign und erlebten so auch die Wirkung unterschiedlicher Übungsvarianten. Dort ließ sich durch einen direkten Vergleich der Übungsmethoden miteinander eine ganz klare Präferenz für das MC-Testen mit Feedback gegenüber der hier verwandten Study-only-Version nachweisen. Die subjektive

Bewertung ohne konkrete Vergleichsperspektive hinterlässt vermutlich nur bei sehr deutlich ins Auge springenden Methodenunterschieden eine nachweisbare Wirkung, was sich hier in der höheren Anstrengung bei massierten Übungen relativ klar abzeichnete.

Studieren und Testen zu Hause in der natürlichen Lernumwelt

Die Durchführung der gesamten Untersuchung wurde via Internet organisiert. Dennoch heben sich die Studien erheblich von solchen Internetexperimenten ab, die jedem beliebigen Internetsurfer eine Teilnahme ermöglichen. Die Probanden wurden vor dem Experiment mehrfach über das Prozedere instruiert. Ihnen wurde ein Plan ausgehändigt, der die korrekte Abfolge der verschiedenen Sitzungen unterstützen sollte. Die einzelnen Übungen waren nur über Passwort erreichbar. Das Computerprogramm legte Ablauf und Zeit der Übungen strikt fest. Nicht kontrollierbar war lediglich die Art des Lernens, was aber durchaus beabsichtigt war, weil die Studierenden ja so üben sollten, wie sie dies normalerweise für sinnvoll und lernförderlich hielten. Deshalb vermute ich nicht ohne Grund, dass die erhobenen Daten während der Übungen ziemlich verlässliche Informationen beinhalten, was etwa die sehr plausiblen und konsistenten Daten zu den Studier- und Feedbackzeiten für richtig und falsch gelöste Items nahe legen. Kritisch einzuschätzen bleibt lediglich die unkontrollierte Erfassung des Behaltens im Posttest. Auch wenn wegen der strikten Anonymität keinerlei "objektiv erkennbarer" extrinsischer Nutzen einer Datenverfälschung erkennbar ist, so bleibt nicht ausgeschlossen, dass einige Studierende "ein möglichst gutes Ergebnis" auch um den Preis unlauterer Hilfen angestrebt haben könnten. Derartige Bedenken hätte man vollständig zerstreuen können, wenn wenigstens die Behaltentestung unter streng kontrollierten Bedingungen im Labor durchgeführt worden wäre. Trotz dieser Einschränkungen liefern die Behaltenswerte des Experimentes 2 meiner Meinung nach eine verwertbare Datenbasis für das Lernen in der natürlichen Lernumgebung.

Erneut hat sich eine riesige Personenstreuung in den Behaltenswerten gezeigt, die versuchsplantechnisch betrachtet aber lediglich Fehlervarianz repräsentiert und die Entdeckung von Methodeneffekten massiv erschwert. Aber so sieht es eben in der pädagogischen Wirklichkeit aus. Man sollte sich stets die Frage stellen, was von den signifikanten Effekten hoch kontrollierter Laborexperimente mit teilweise artifiziellen Lehrinhalten nach der Übertragung im normalen Unterricht überhaupt noch ankommt.

Stabiles Wissen ohne Wiederholung schwer vorstellbar

Als anspruchsloses Lehrziel genießt Faktenwissen bei Pädagogen keinen guten Ruf, obgleich es häufig im Schulsystem vermittelt und in Prüfungen verlangt wird. Ein Teil des Faktenwissens ist zwingend notwendig, da es für höhere Lernprozesse wie Anwendung und Problemlösung zwingend verfügbar sein muss. Dies gilt jedoch nicht für das Erlernen der Bundesstaaten der USA. Die Studien verfolgten in erster Linie das Ziel, einige pädagogische Instruktionsprinzipien des Übens zu überprüfen, was mit der Hoffnung verbunden war, es an einfachsten Lehrzielen noch am ehesten nachweisen zu können. Unmittelbar nach dem Erwerb des Faktenwissens setzt das Vergessen ein, sofern das Wissen nicht durch Wiederholungen aufgefrischt wird. Obgleich diese Studie das Behalten nach einem Retentionsintervall von einer Woche erfasste, was für Forschungszwecke schon recht lange erscheint, ist dieses Zeitintervall gemessen am pädagogischen Anspruch, stabiles Wissen vermitteln zu wollen, verschwindend gering. Carpenter, Pashler & Cepeda (2008), welche die Wirkung von Übungen beim Er-

lernen historischer Fakten im Schulsystem analysierten, fanden nach einer Woche ca. 50% und nach 4 Monaten noch ca. 35% korrekte Antworten. Auch nach 9 Monaten konnten noch Unterschiede verschiedener Übungsmethoden statistisch gesichert werden. Unter der Bedingung Testen mit Feedback wussten die Schüler noch ca. 10%, unter der Bedingung erneutes Studieren noch ca. 7 % und bei keiner vorausgehenden Übung 5 % der historischen Fakten. Selbst der Effekt des verteilten Lernens ließ sich noch bestätigen. Die so verbliebenen Methodenunterschiede sind jedoch jenseits aller Signifikanzüberlegungen praktisch ohne ernsthaften Belang. Die Vorstellung, die Schule könne Wissen vermitteln, dass quasi jahrelang konserviert nur darauf wartet, je nach Bedarf verfügbar zu sein, bleibt **nicht nur für Faktenwissen** eine völlig überzogene Illusion. Fast alles, was nicht wenigstens gelegentlich in der einen oder anderen Form wiederholt wird, fällt langfristig dem Vergessen anheim. Es bleibt die Hoffnung, dass das, was schon einmal gründlich eingeprägt und angemessen wiederholt wurde, durch erneute Übung im Bedarfsfalle dem Gedächtnis schneller wieder verfügbar gemacht werden kann.

Literatur

- Bahrick, H. P. & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory & Language*, 52(4), 566-577.
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2008). Using tests to enhance 8th grade students' retention of U. S. history facts. *Applied Cognitive Psychology*. Published Online: 26 Aug 2008.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. *Psychological Bulletin*, 132, 354-380.
- Cull, W. L. (2000). Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall. *Appl. Cognit. Psychol.* 14, 215-235.
- Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1, 309-330.
- Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of Perception and Cognition*. Vol. 10, Memory, pp. 317-344. New York: Academic Press.
- Donovan, J. J. & Radosevich, D. R. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, 84, 795-805.
- Kang, S. H. K., McDermott, K. B. & Roediger, H. L., III. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology* 19, 528-558.
- Kornell, N & Bjork, R.A (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review* 14 (2), 219-224.
- Jacobs, B. (2004). [Lohnt sich Antwort abhängiges Feedback ?](#)
URN: urn:nbn:de:bsz:291-psydok-2130
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2004/213/>
- Jacobs, B. (2006a). [Erneutes Studieren oder Testen mit Feedback beim Einüben von Faktenwissen am Beispiel des Erlernens der Bundesstaaten der USA.](#)
URN: urn:nbn:de:bsz:291-psydok-5992
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2006/599/>
- Jacobs, B. (2006b) [Die Auswirkungen von Short-Answer- und Multiple Choice-Übungsaufgaben auf das Lernergebnis.](#)
URN: urn:nbn:de:bsz:291-psydok-7609
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2006/760/>
- Jacobs, B. (2007). [Die Behaltenswirksamkeit wiederholten Einprägens im Vergleich zu Computer- und selbst gesteuertem Testen mit Feedback.](#)
<http://www.phil.uni-sb.de/~jakobs/wwwartikel/teststudy/teststudy2.html>

- Jacobs, B. (2008a). Führt selbst gesteuertes Testen mit Feedback zu höheren Behaltensleistungen als das Einprägen mit Hilfe einer Landkarte?
<http://www.phil.uni-sb.de/~jakobs/wwwartikel/teststudy/teststudy3.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. (2008c). 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/>
- LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, 67, 259-266.
- Pashler, H., Rohrer, D., Cepeda, N. J., Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review* 2007, 14 (2), 187-193.
- Pavlik, P.I. & Anderson, J. R. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cognitive Science* (29), 559-586.
- Rohrer, D. & Pashler, H. (2007). Increasing Retention without Increasing Study Time. *Current Directions in Psychological Science*. 16 (4), 183-186.
- Weigold, A. (2008). The Relationship Between Restudying and Testing in the Short and Long Term. Dissertation. Texas Tech University
http://etd.lib.ttu.edu/theses/available/etd-03142008-140331/unrestricted/Weigold_Arne_Diss.pdf [20.6.2008]