Autor: Bernhard Jacobs, Februar 2006
Medienzentrum der Philosophischen Fakultäten der Universität des Saarlandes

Erneutes Studieren oder Testen mit Feedback beim Einüben von Faktenwissen am Beispiel des Erlernens der Bundesstaaten der USA.

Zusammenfassung

Zentrale Aussagen zur Lernwirksamkeit von Aufgabenstellungen und Feedback lauten, bereits das Testen allein stelle eine wirksame Übung dar und Testen mit Feedback sei dem wiederholten Studieren hinsichtlich des längerfristigen Behaltens überlegen. Diese schon häufiger bestätigten Hypothesen sollten an einem bisher selten untersuchten Lerninhalt, dem Zuordnen von Namen zu geografischen Gebilden, erneut einer Prüfung unterzogen werden. Zu diesem Zweck wurde die Wirkung wiederholten Einprägens und drei verschiedener Test- bzw. Feedbackvarianten auf das Behalten getestet und gewisse empirische Evidenzen für die Bestätigung der Hypothesen gefunden. Aufwändige formale Test - und Feedbackprozeduren mit Flashcard bzw. Antwort abhängigem Feedback schneiden nicht besser ab als eine einfache informelle Testform, welche lediglich ein Nachdenken bzw. Erinnern einfordert, wonach dann die Rückmeldung der korrekten Antwort (Knowledge of correct response) folgt. Der klassische Short Answer-Aufgabentyp deutet eine leichte Überlegenheit gegenüber dem MC-Aufgabentyp an, erfordert aber deutlich mehr Übungszeit. Subjektive Einschätzungen der Lernwirksamkeit der 4 Übungsmethoden stimmen nur teilweise mit den objektiven Behaltensdaten überein.

Schlagworte: Aufgabenformen, Übung, Practice, Feedback, Feedbackarten, Drill, Multiple Choice, Short Answer, Flashcard

Gliederung

Problemstellung und Zielsetzung

Der gemeinhin geringen pädagogischen Wertschätzung des Faktenwissens steht seine Notwendigkeit als Grundlage für anspruchsvollere Lernziele gegenüber. Nur wenn bestimmte Fakten dem Gedächtnis schnell und zuverlässig zur Verfügung stehen, lassen sich übergeordnete Lernprozesse wie Begriffswissen, Regellernen sowie Problemlösung effizient anregen. Die neuen Medien bieten vielfältige Möglichkeiten Faktenwissen zu fördern und Übungen anzubieten, um die Stabilität des Behaltens zu verbessern. Hierbei soll insbesondere über Möglichkeiten nachgedacht werden, welche im Gegensatz zu einer Papierfassung relativ bequem eine Testung mit unmittelbarem Feedback anbieten.

Techniken zum Erlernen neuen Wissens lassen sich grob vereinfacht klassifizieren in

- Organisationstechniken,
- Elaborationstechniken,
- Wiederholungstechniken.

Im Mittelpunkt stehen hier simple Wiederholungstechniken (Practice) des erneuten Einübens, bei der lediglich die Ausgangsinformationen mehrmals zur Bearbeitung im Sinne eines Restudy's oder Testens mit Feedback vorgelegt werden. Wenngleich sich zum Erlernen von Faktenwissen Wiederholungstechniken geradezu anbieten, da Faktenwissen kein Verständnis voraussetzt, die Rückmeldung der korrekten Antwort hinreichend erscheint (Jacobs 2002) und Transfer gar nicht zur Disposition steht, werden die auch für Faktenwissen förderlichen Wirkungen anderer Lernstrategien damit keineswegs in Frage gestellt, sondern lediglich nicht behandelt. Die Informationsdarbietung des Computers in Form eines Drills erzwingt keineswegs geistlos mechanisches Auswendiglernen (rote learning). Wie der Lerner selbst das Enkodieren oder Ins-Gedächtnisrufen vornimmt, bleibt ihm vielmehr selbst überlassen und ist nicht Gegenstand dieser Arbeit.

Als umfassendes Lehrziel dient die korrekte Zuordnung der Namen der US-Bundesstaaten zu ihren Territorien auf der Landkarte. Der Lerner soll hierbei zum einen dem geographischen Gebilde eines Staatsgebietes auf der Landkarte der USA den entsprechenden Staatsnamen zuordnen, zum andern bei Vorgabe des Namens das entsprechende Staatsterritorium lokalisieren können. Hierbei handelt es sich um eine besondere Art von Paarassoziationslernen, die immer erforderlich ist, wenn Bezeichnungen räumlich zugeordnet werden müssen, wie etwa auch beim Erlernen der Bestandteile eines Skelettes oder dem Einprägen von Einzelteilen eines technischen Gerätes bzw. einer Zeichnung desselben.

Zentrale, schon öfter bestätigte Hypothesen, das reine Testen stärke das Behalten sowie Testen mit Feedback erziele einen stabileren Behaltensgewinn als wiederholtes Einprägen, sollen erneut einer empirischen Prüfung unterzogen werden. Darüber hinaus wird die Wirksamkeit etlicher Varianten getestet, die in unterschiedlicher Weise Erinnerungsprozesse anregen oder die Testung gestalten. Der Multiple Choice Aufgabentyp verlangt bei Faktenwissen eher ein Wiedererkennen, während der Short Answer-Aufgabentyp [Deutsch: Kurze Freiantwortaufgabe; englisch auch: constructed response recall study task] eine freie Wiedergabe einfordert, die möglicherweise einen wirksameren Gedächtnisabruf nach sich zieht. Ist eine formale Testung, welche die explizite Antwort des Lerners mit "richtig bzw. falsch (=Knowledge Of Result (KOR)) bewertet, notwendig oder reicht eine informelle, selbst kontrollierte Überprüfung mittels Know-

ledge of Correct Response (KCR) aus? Schließlich soll die Auswirkung einer verstärkten Bearbeitung fehlerhafter Items eingeschätzt werden.

Bisherige Forschung zur Wirksamkeit des Testens und dem Vergleich erneuten Studierens vs. Testens

Eine zentrale didaktische Frage jeder Lernorganisation lautet, wann von einer Informationsdarbietung zu einer Testung übergegangen werden soll. Zunächst muss entschieden werden, wie gründlich der Lehrstoff in einer Lernaneignungsphase vermittelt wird. Danach ist abzuwägen, ob die Information zwecks weiterer Stabilisierung wiederholt zum Enkodieren dargeboten oder besser direkt getestet werden soll. Zwischen wiederholter Präsentation und Testung sind natürlich Übergangs- bzw. Zwischenformen denkbar - etwa Variationen der Präsentation bzw. Testen mit Hilfestellung - auf die hier jedoch nicht weiter eingegangen werden soll. Die Entscheidung hinsichtlich der Frage, Wann macht es Sinn, Aufgaben zu stellen? (Jacobs 2003) hängt weiterhin vom Lehrzielniveau sowie dem Schwierigkeitsgrad ab und hierbei wird angenommen, bei reinem Faktenwissen könne in einem relativ früheren Lernstadium zum Testen (mit Feedback) übergegangen werden.

Experimentelle Studien im pädagogischen Umfeld erbrachten den Nachweis, die Durchführung eines Tests im Anschluss an eine Instruktion fördere das langfristige Behalten (z.B. (Duchastel & Nungester (1982), Glover (1989), Haynie (1994)). Die Metaanalyse von Hamaker (1986) zur adjunct question Forschung siedelt den Lernerfolg von Fragen bzw. Tests (ohne Rückmeldungen) bei einfachem Faktenwissen in einer Größenordnung von ca. einer Effektstärke an.

Etliche psychologische Experimente belegen einen Behaltensvorteil reinen Testens gegenüber dem erneuten Studieren beim Textlernen (Nungester & Duchastel (1982), Dempster, Dunbar & Corkill (2001), Richland et al.(2005), Listen- oder Paarassoziationslernen (etwa; Carpenter & DeLosh (2005); siehe zusammenfassend: Jacobs 2005a) sowie beim Free-recall-lernen, (siehe zusammenfassend Jacobs 2005b). Testen wird hierbei mit einer ernst zu nehmenden Kontrollbedingung, etwa erneutem Einprägen verglichen. Der Behaltensvorteil des Testens zeigt sich vornehmlich nach einem längeren Retentionsintervall. Testen verbessert somit die Behaltensstabilität. Der überraschende Vorteil des reinen Testens (ohne Feedback) gegenüber der erneuten Informationsdarbietung ist jedoch nicht immer zu erwarten, sondern geht offenbar auf die besondere Versuchsanordnung zurück. Die Experimente sind meist so aufgebaut, dass die Testung zu sehr hohen Erfolgsquoten führt (z.B. Kuo & Hirshman (1996)), mithin einen recht erfolgreichen Abruf aus dem Gedächtnis ermöglicht. Weitere Testungen können die Behaltensleistung aber nicht mehr erhöhen, sondern lediglich das Vergessen mindern (z.B. Roediger & Karpicke (2005), wenn man einmal solche exotischen Phänomene wie Hypermnesia (Wheeler & Roediger. (1992)) ausklammert, die Cull (2003, S. 229) bei keinem seiner 4 Experimente bestätigen konnte. Mehrere Experimente von Carpenter & DeLosh (2005) zum Face-Name Learning belegen eindrucksvoll den Behaltensvorteil des Testens im Vergleich zum erneuten Studieren.

Bei den meisten mir bekannten Studien zur Wirkung des reinen Testens lag die Erfolgswahrscheinlichkeit während der Testung in der Übungsphase deutlich über 50%. **Es macht im Hinblick auf die Förderung einer Lernstabilisierung offenbar wenig Sinn, Lehrgebiete zu testen, die etwa eine Erfolgswahrscheinlichkeit deutlich unter 50% erwarten lassen.** Je geringer das Wissen, desto eher dürfte erneute In-

formationsdarbietung günstigere Lerneffekte als reine Testung bewirken, zumindest bei trivialem Faktenwissen, welches lediglich Einprägung und kein Verständnis erfordert.

Die Studien reflektieren auch keineswegs empirische Eindeutigkeit hinsichtlich der Überlegenheit des Testens gegenüber erneutem Studieren. Cull (2000, S. 216) findet in seinen eigenen Studien unter bestimmten Bedingungen - z..B. sehr gute Lernaneignung, massiertes Üben und langes Rententionsinterval - eine Überlegenheit des reinen Testens im Vergleich zum erneuten Studieren, belegt in den Experimenten 2 und 3 (2000,S 219 table 1) zum Teil aber auch ganz deutlich die Unterlegenheit des Testens gegenüber dem erneuten Studieren, z. B. bei weniger guter Lernaneignung oder verteiltem Lernen.

Testen mit nachfolgendem Feedback verbindet den Vorteil des stabilisierenden Effektes einer erfolgreichen Testung mit den Vorteilen der erneuten Enkodierung im Falle eines Fehlers. Mehrere Experimente belegen den Behaltensvorteil von Fragestellungen mit Feedback im Vergleich zu Restudy. (z.B.: Dempster ,F.N., Dunbar, M.E., Corkill (2001) ,Carrier, M., & Pashler, H. (1992), , Pashler, Cepeda, Wixted & Rohrer (2005), , Morris, Fritz, Jackson, Nichol & Roberts. (2005). Besonders hervorzuheben sind die Experimente von Cull (2000) zum Vokabellernen, die fast ausnahmslos für den Behaltensvorteil von Testen mit Feedback gegenüber erneutem Einprägen sprechen. Clifton (2005) konnte die Überlegenheit des Testens mit Feedback gegenüber gezieltem Studieren sogar für das Lehrziel Verständnis im praktischen Universitätsbetrieb bestätigen. Jacobs (2001) überprüfte bei Kombinatorikproblemen die Bearbeitung ausschließlicher Lösungsbeispiele mit einer Variante, die später Aufgaben mit knappen Feedback verlangte und fand einen höheren Lernerfolg für die getestete Gruppe mit Feedback.

Das Problem der Vergleichbarkeit der Trainingsbedingungen

Um die Auswirkung einer Übungsmethode auf das Lernergebnis abschätzen zu können, müssen sonstige lernfördernde Bedingungen konstant gehalten werden. Insbesondere ist eine Konfundierung von Übungszeit und Lernmethode zu vermeiden.

Es erscheint auf den ersten Blick plausibel anzunehmen, Trainingsmethode A sei Trainingsmethode B überlegen, wenn sie bei vergleichbarer Trainingszeit einen höheren bzw. stabileren Lerngewinn erzielt. Eine Alternative bestünde darin, Wirksamkeitsunterschiede in der Zeit zu messen, die notwendig ist, um ein bestimmtes Leistungskriterium zu erzielen. Da das aktuelle Erreichen eines Leistungsniveaus allerdings nicht mit langfristigem Behalten verwechselt werden darf, greift diese Methode nur für den Nachweis des unmittelbaren Lernerfolgs und erscheint problematisch für Fragestellungen, welche etwa eine begründete Interaktion zwischen Trainingsmethoden und Behaltensintervallen erwarten lassen (siehe z.B. Schmidt & Bjork (1992), Wheeler, Ewers & Buonanno 2003, Roediger & Karpicke (2005), Rawson & Kintsch (2005), Richland et. al. (2005). Schließlich könnte man den Kriteriumserfolg als diejenige Zeit betrachten, welche für einen bestimmten langfristigen Lernerfolgszuwachs notwendig ist, was dann aber immer nur im Nachhinein festgestellt und schwerlich experimentell hergestellt werden könnte.

Ein fairer Vergleich von Übungsmethoden erfordert gleiche Lernzeiten. Deren Ausgestaltung kann aber unterschiedlich realisiert werden. Im Experiment von Carrier & Pashler (1992) wurde das Zeitproblem durch eine vom Computer starr vorgegebene vergleichbare Studier- bzw. Testzeit pro Aufgabe gelöst. Für das Studium eines Vokabelpaares waren 10 Sekunden vorgesehen, beim Testen sah der Lerner 5 Sekunden den Cue (z.B. Eskimo Wort) und nach weiteren 5 Sekunden zusätzlich noch das target (=KCR-Feedback =englisches Wort). Für etliche Lerner bzw. Aufgaben könnte die

lange Präsentationsdauer aber auch als unnötig verschenkte Lernzeit angesehen werden, welche das erneute Studieren gegenüber dem Testen benachteiligt. So fanden etwa Bahrick und Hall (2005) heraus, dass Studenten bei frei bestimmbarer Bearbeitungszeit schwierige Vokabeln mit ca. 8 Sekunden deutlich länger einprägten als leichte Vokabeln mit ca. 2 Sekunden.

Hier sollte der Lerner in der Übungsphase aus Gründen der ökologischen Validität sein Lerntempo selbst bestimmen, was unweigerlich interindividuelle Variationen der Übungsmenge nach sich zieht. Die Qualität einer Lernmethode zeichnet sich aber auch dadurch aus, bei konstanter Übungszeit einen größeren Umfang von Übungsmöglichkeiten zu gewähren. Entscheidend ist letztendlich der Lernzuwachs in einer Zeiteinheit. Jede Trainingsvariante sollte eine solide Grundlage für einen unmittelbar hohen und mindestens mäßig stabilen Lernerfolg nach sich ziehen. Leider war es aus praktisch organisatorischen Gründe nicht möglich, das Training in Form einer verteilten Übung mit einem längeren Zeitabstand auf mindestens 2 Termine zu verlegen, um den spacing Effekt besser auszunutzen. (Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (in press), Bahrick und Mitarbeiter, z.B. Bahrick und Hall (2005).

Man darf Zweifel daran hegen, ob der Zusammenhang zwischen Lernzeit und Lernzuwachs für alle Varianten gleich ausfällt, eine Annahme, die man stillschweigend bei der Forderung gleicher Lernzeiten einfach unterstellt. Hierbei stimmen vor allem empirische Befunde bedenklich, die aufzeigen, dass wiederholtes Lesen die Lernleistung oftmals nur marginal steigert. (etwa Karpicke und Roediger (2005)) Demnach wäre es durchaus möglich, dass manche Trainingsvarianten auch bei weniger Lernzeit denselben Lernerfolg nach sich zögen und vergleichbare Lernzeiten eben nicht vergleichbares Lernpotential bedeuten. Letztlich benötigte man Daten zum Lernzuwachs im Zeitverlauf (etwa: Lernerfolg = f (Übungszeit (Übungsmethode))).

Wiederholtes Einprägen und verschiedene Test bzw. Feedbackvarianten

Wenngleich im Anschluss an eine Lernaneignungsphase das Testen gegenüber der erneuten Präsentation des Lehrstoffs häufig Lernvorteile erbrachte, so wäre es pädagogisch unbefriedend, beide Lerntechniken als sich gegenseitig ausschließende Alternativen zu betrachten, weil ganz offensichtlich die Positionierung jeder Maßnahme an der richtigen Stelle entscheidende Vorteile verspricht. Ein sich der Testung eines Items anschließendes informatives Feedback ist nicht anderes als das Angebot einer erneuten Informationsaufnahme, die sich allerdings gezielt auf die gestellte Frage bezieht und insbesondere bei Fehlern wertvolle Korrekturhilfen anbietet. Es erscheint somit nützlich, Informationen vornehmlich dann anzubieten, wenn Defizite erkennbar sind. Durch das Testen werden zum einen Erinnerungsbemühungen angeregt, welche das Behalten im Falle der korrekten Antwort verbessern, zum andern werden Schwächen diagnostiziert, welche einen höheren Informationsbedarf signalisieren. Deshalb galt für die Konstruktion der experimentellen Bedingungen die Zielrichtung, den Lerner insbesondere bei Fehlern soweit wie möglich zu unterstützen und bei der erneuten Präsentation von Information besonderen Wert auf die Überwindung der Fehler zu legen.

Erneutes Darbieten bzw. Enkodieren der Fakteninformation geht normalerweise aber deutlich zügiger von statten als formelles Testen. Da die Lernzeit beider Übungsmethoden jedoch konstant zu halten war, wurde zunächst nach einer Methode Ausschau gehalten, den Ablauf der Testung zu beschleunigen. Um einen möglichst ausführlichen Retrieval anzuregen, sollte die Testung durch den sonst sehr aufwändigen Short Answer Aufgabentyp auf möglichst ökonomische Art und Weise durchgeführt werden. Für Übungszwecke lässt sich der Short Answer-Aufgabentyp einfacher gestalten, wenn man die zu leistende geistige Anforderung eines Abrufs aus dem Gedächtnis formal

einfordert, auf eine schriftliche Eingabe und damit eine formale Testung jedoch verzichtet und auf Verlangen Knowledge of Correct Response Feedback (KCR) gewährt. Dieser verdeckte Short Answer Typ erwartet nach der Fragestellung eine Antwort im Geiste und überlässt die Auswertung dem Lerner (= KCR ohne formelles Knowledge Of Result (KOR, z.B. richtig oder falsch).

Short Answer- und MC-Test sind demgegenüber zwingende Testverfahren, die auch eindeutige Rückmeldungen zur Korrektheit der Antwort zulassen, wenngleich auch diese Methoden durch Eingabe beliebiger Antworten intentionswidrig umgangen werden können. Nur bei einer echten Testung kann der Computer die diagnostischen Möglichkeiten für weitere Verbesserungen ausnutzen. Als Rückmeldungen folgen stets KOR, KCR sowie die potentielle Verwechslung, die teilweise auch als Antwort abhängiges Feedback bzw. Response Contingent Feedback (RCF, siehe Jacobs (2004)) gedeutet werden kann und z.B. auch bei Siegel & Misselt (1984) zur Anwendung kam. Klickt der Proband z.B. bei der Aufforderung "Indiana" auf das Gebiet von "Ohio", so erscheint im Bereich der Maus "Ohio" (= response contingent feedback) und zusätzlich wird "Indiana" (=Knowledge of Correct Response, KCR) an der zutreffenden Stelle platziert. Zudem erhielten beide echten Testvarianten eine Flashcard-Bedingung, welche falsch beantwortete Fragen sooft jeweils an das Ende der Aufgabenliste verlagerte, bis alle Aufgaben einmal korrekt beantwortet waren.

Durch die folgende empirische Studie sollten im Wesentlichen geprüft werden,

- ob zusätzliches Testen langfristiges Behalten stärkt,
- Testen mit Feedback gegenüber erneutem Studieren Behaltensvorteile erbringt, sowie
- welche Test- bzw. Feedbackprozedur am besten für Übungszwecke geeignet ist.

Die empirische Studie

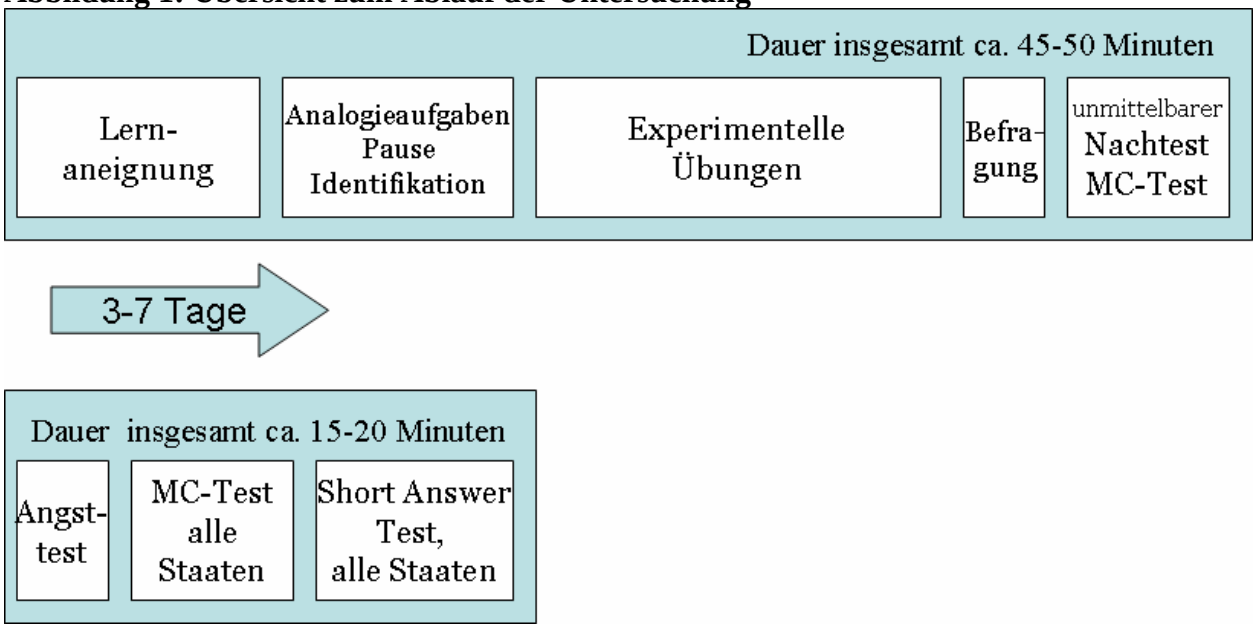
Versuchspersonen

An der Untersuchung nahmen 5 Bedienstete sowie 61 Studierende der Universität des Saarlandes teil. Das Durchschnittsalter betrug 22 Jahre. Lehramtsstudierende des vom Autor geleiteten Seminars "Gestaltung von Lernmaterialien" sowie Studierende der Informationswissenschaften, die ein Tutorium absolvierten, wurden im Rahmen der jeweiligen Lehrveranstaltung zur Teilnahme am Experiment motiviert und absolvierten das Lernexperiment in den Cip-Räumen der Philosophischen Fakultäten der Universität des Saarlandes. An dieser Stelle gilt mein Dank den Tutoren Herrn Wadle und Frau Burgard für die Organisation von Probanden, sowie allen beteiligten Probanden, die etwas mehr als 1 Stunde ihrer Zeit unentgeltlich der Wissenschaft zur Verfügung stellten.

Den Probanden war vor dem Experiment klar gemacht worden, dass ein zweiter Untersuchungstermin anstehe, jedoch keine Angaben darüber gemacht, was dort erfasst werden sollte. Den Nachtest, 3 bis 7 Tage später, bearbeiteten die Probanden entweder in den Cip-Räumen oder von zu Hause aus via Internet. 57 Probanden haben den Nachtest wahrgenommen.

Untersuchungsablauf

Abbildung 1: Übersicht zum Ablauf der Untersuchung



Die in Abbildung 1 verdeutlichten Phasen des Experimentes werden nachfolgend genauer spezifiziert und begründet.

Lernaneignung und Übungsphase

Zunächst sahen die Lerner zum Überblick für eine halbe Minute alle Staaten der USA auf der Landkarte, wobei der Staatsname innerhalb des entsprechenden Territoriums platziert war. Anschließend wurden die 50 Staatsnamen nacheinander in einer nachvollziehbaren Reihenfolge einzeln für jeweils 5 Sekunden dargeboten: Dieser Ablauf gestaltete sich wie folgt: Staatsname 1 erscheint im betreffenden Staatsgebiet 1 für 5 Sekunden. Dann verschwindet der alte Staatsname 1 und der nachfolgende Staatsname 2 erscheint im zutreffenden Gebiet 2 für 5 Sekunden...usw. Nachdem so alle Staaten dargeboten wurden, folgte ein weiterer Darbietungsdurchgang auf einem anderen, ebenfalls nachvollziehbaren Wege. Insgesamt betrug die reine Darbietungszeit in der Lernphase 10,5 Minuten. In der Lernaneignungsphase sollten die Probanden zunächst die Bundesländer der USA kennen lernen. Es war beabsichtigt, die Lernaneignungsphase so auszurichten, dass etwa 50 % aller Staaten korrekt erkannt werden konnten. Durch reine Informationsdarbietung sollten gewisse Grundlagen geschaffen werden, die aber noch Spielraum für die nachfolgenden experimentellen Übungen erlauben mussten.

Füllaufgabe

Um die Massivität der gesamten Übung etwas zugunsten von mehr verteiltem Lernen zu durchbrechen, bearbeiteten die Probanden nach der Lernaneignung 10 Aufgaben eines Analogietests, erhielten anschließend Rückmeldung über ihr Ergebnis, mussten dann eine Minute pausieren und nachfolgend Angaben zur anonymen Identifizierung machen. Vor der eigentlichen experimentellen Übungsphase wurden sie via Computer über den Sinn und Zweck umfangreicher Übungen informiert und zu einem zügigen, aber konzentrierten Arbeiten motiviert. Der Zeitabstand zwischen dem Ende der Lernaneignung und dem Beginn der experimentellen Übungsphase betrug insgesamt ca. 8 bis 10 Minuten.

Die experimentellen Übungsmethoden

Die experimentellen Bedingungen wurden in einem Wiederholungsdesign realisiert. Dabei wurden nicht alle Staaten der USA, sondern jeweils ein Staatenbereich bzw. eine Gruppe von ca. 12 Staaten mit einer Methode eingeübt. In Anlehnung an die überwiegend angloamerikanische Feedbackliteratur verwende ich für die einzelnen Treatments englische Namen

Tabelle 1: Beschreibung der Übungsmethoden und Demonstrationsbeispiele

Übungsmethode , Login-Informationen: Username: Alabama, Passwort: Alabama	Demonstrations-beispiel*
Erneutes Studieren Study only (**SO):** Ein Staatsname erscheint im dazugehörigen Staatsgebiet. Der Lerner bestimmt durch Tastendruck die Auswahl des nächsten Staates. Nach jedem Durchgang sieht der Lerner für 20 Sekunden alle Staatsnamen eines eingeübten Staatsbereiches.	**Erneutes Studieren** für Staatsbereich 1
Covert Short Answer mit KCR-Feedback (CSA): Ein Fragezeichen erscheint in einem Staatsgebiet und verlangt ein Erinnern des entsprechenden Staatsnamens. Der Lerner bestätigt die gedachte Antwort durch Mausklick oder durch Tippen auf die Leertaste. Daraufhin erscheint an der Stelle des Fragezeichens der Staatsname als Rückmeldung KCR. Nach jedem Durchgang sieht der Lerner für 20 Sekunden alle Staatsnamen eines Staatsbereiches.	**Covert Short Answer mit KCR-Feedback:** für Staatsbereich 2
Multiple Choice Test mit KOR+KCR+ partiellem RCF Feedback+ Flashcard: (MC) Ein Staatsname wird verbal vorgegeben. Sein Gebiet soll mit der Maus angeklickt werden. Anschließend folgt symbolisch KOR (richtig/falsch). Zusätzlich erscheint der zutreffende Staatsname (KCR), bei einer Verwechslung zusätzlich der falsch angeklickte Staat im entsprechenden Staatsgebiet (RCF). Nach Beendigung jedes Flashcard-Durchgangs sieht der Lerner für 12 Sekunden alle zuvor falsch beantworteten oder verwechselten Staaten der eingeübten Staatengruppe.	**Multiple Choice Test mit KOR+KCR+ partiellem RCF Feedback+ Flashcard:** für Staatsbereich 3
Short Answer mit KOR+KCR+ partiellem RCF Feedback+ Flashcard: (SA) Ein Fragezeichen erscheint in einem Staatsgebiet und verlangt das Eintippen des entsprechenden Staatsnamens in exakter Schreibweise. Anschließend folgt symbolisch KOR (richtig/falsch). Zusätzlich erscheint der zutreffende Staatsname (KCR), bei einer Verwechslung der Name des falschen Staates im entsprechenden Staatsgebiet. Nach Beendigung jedes Flashcard-Durchgangs sieht der Lerner für 12 Sekunden alle zuvor falsch beantworteten oder verwechselte Staaten.	**Short Answer mit KOR+KCR+ partiellem RCF Feedback+ Flashcard:** für Staatsbereich 4

* Sollten die WWW-Seiten der experimentellen Methoden nicht erreichbar sein, bitte an b.jacobs@mx.uni-saarland.de wenden
***** Der echte Versuch lief im Vollbildmodus ohne Menu-, Symbol-, oder Statuszeilen.
Die Programme dienen nur als Demonstration. Das JavaScript-programm ist unverständlich und beinhaltet etliche Eigentümlichkeiten.

Gemeinsamkeiten und Unterschiede der Übungsmethoden

Beim erneuten Studieren platziert der Computer zunächst die Namen der Bundesstaaten analog der Lernaneignungsphase einzeln nacheinander in die entsprechenden Gebiete auf der Landkarte. Im Gegensatz zur Lernaneignungsphase war die Reihenfolge der Staaten für jede Vp nach Zufall bestimmt worden. Diese Präsentationsfolge erschwert eine implizite Testung des Lerners, wie Sie bei einem geordneten Weg durch

die Staaten möglich wäre (z.B.. " Nach California muesste jetzt eigentlich Nevada folgen.... tatsächlich...) Die Gruppe erneutes Studieren unterliegt einer Reread- bzw. Study only -oder Restudy Bedingung, da ihr lediglich die Information mehrmals zum Einprägen präsentiert wird. Sie fungiert als entscheidender Vergleichmaßstab zu den restlichen Übungsvarianten, welche alle dazu auffordern, Erinnerungsprozesse (retrieval) in Gang zu setzen.

Erneutes Studieren und verdecktes Short Answer mit KCR-Feedback unterscheiden sich hier lediglich durch die der Informationsdarbietung vorausgehenden Fragestellung beim Testen. Das Testen pro Aufgabe erfordert lediglich einen Klick mehr auf die Leertaste und ist somit vom Zeitaufwand äußerst ökonomisch. Es bleibt schwer abzuschätzen, ob und wie ernsthaft der Lerner Erinnerungsbemühungen in Gang setzt, bevor er die korrekte Antwort anfordert. Es besteht eine gewisse Missgebrauchsgefahr, das Testangebot zu ignorieren und direkt durch Doppelklick auf die Leertaste die korrekten Antworten einzusehen. Bei sinnvollem Einsatz verspricht die Methode allerdings hohen Lernerfolg in sehr kurzer Zeit und könnte für Aufgabenstellungen mit direkt verständlichen Rückmeldungen eine echte Alternative zu formaler Testung darstellen.

Short Answer und Covert Short Answer unterscheiden sich durch die Antwortart (Eintippen vs. Denken), die formalen Rückmeldungen (bei Covert SA nur KCR) sowie die Organisation der Itemwiederholungen (Flashcard mehr falsche Items vs. gleiche Anzahl für alle Items) und die Gesamtsicht der Staaten (falsche und verwechselte Staaten vs. alle Staaten). Da nur bei echter Testung in Erfahrung gebracht werden konnte, welche Staaten richtig oder falsch beantwortet wurden, ließ sich die Gesamtsicht nur bei den echten Testverfahren auf die falsch gelösten Staaten begrenzen. Ziel war es ja, besonders die schwierigen Items verstärkt zum Enkodieren zu präsentieren. Die etwas kürzere Zeit für die Gesamtsicht bei den echten Testvarianten lässt sich mit der geringeren Staatenanzahl begründen.

Experimenteller Aufbau der Übungsitems

Die riesige Personenvarianz in einem Vorversuch bei gleichzeitig beschränkten Zugriffsmöglichkeiten auf Probanden zerstörte die ursprüngliche Planung, die Übungsmethoden in einem vollständig randomisierten Versuchplan mit unabhängigen Gruppen zu prüfen und verlangte nach einem Wiederholungsexperiment, um überhaupt eine Chance, Effekte nachzuweisen, wahrnehmen zu können.

Die Bundesstaaten der USA (Items) wurden deshalb in 4 Bereiche (Gruppen, Zonen) eingeteilt, die 12 bzw. 13 Staaten umfassten. Jeder der 4 Staatsbereiche wurde nacheinander mit einer anderen Methode eingeübt. Die Reihenfolge der Staatsbereiche und die Reihenfolge der experimentellen Methoden wurden dabei ausbalanciert. Für jede spezielle Versuchskonstellation wurden 2 Fassungen gebildet, denen 2 Probanden zugeordnet wurden. Dann wurde durch Zufall bestimmt, welcher dieser beiden Probanden den unmittelbaren Nachtest erhält. (genaues Vorgehen)

Die Zeitdauer für die einzelnen Übungsvarianten wurde bis auf eine Ausnahme auf exakt 4 Minuten festgelegt. Dies entspricht einer durchschnittlichen Gesamtübungszeit von ca. 30 Sekunden für ein Item (10 in der Lernaneignungsphase und 20 in der experimentellen Übung). Die durchschnittliche Gesamtbearbeitungszeit von 30 Sekunden pro Item war intuitiv offenbar gut gewählt. Denn Untersuchungen von Rohrer et al. (2005, S. 371) zufolge, stärkt eine Übungszeit von 30 Sekunden pro Item ziemlich op-

timal das längerfristige Behalten für ein Retentionsintervall von einer Woche. Bei der besonders Zeit intensiven Short-Answer-Aufgabenform bestand die Gefahr, dass etliche Lerner zu wenige Items innerhalb der zulässigen Zeit bearbeiten würden, womit das Flashcard-Test-Konzept nicht hätte realisiert werden können. Deshalb wurde bei der Short-Answer-Test-Übung mit Feedback auf die Durchführung eines vollständigen Flashcard-Übungsdurchgangs bestanden und notfalls die Übungszeit über 4 Minuten ausgedehnt. Die Übungszeit von 4 Minuten wurde mit der Intention gewählt, dem durchschnittlichen Lerner in der Regel bei allen Übungsvarianten außer Short Answer mindestens 2 Lerndurchgänge zu ermöglichen.

Abhängige Variablen

Während der Übung wurden einige Prozessvariablen, wie z.B. die Anzahl der bearbeiteten Items erhoben. Nach der Übungsphase folgte eine kleine Befragung zur Evaluation. Abschließend bearbeitete ein Teil der Probanden den unmittelbaren MC-Nachtest. Hierbei wurde ein Staatsname schriftlich vorgeben, dessen Staatsgebiet mit der Maus angeklickt werden sollte. Bei diesem technisch auch unter dem Namen clickable map bezeichneten Aufgabentyp handelt sich nach Rütter (1973) um ein Multiple Choice Format mit 50 Alternativen, weil die korrekte Antwort in der Aufgabenstellung enthalten ist. Alle 50 Staaten wurden für jede Vp in zufälliger Reihenfolge zur Testbearbeitung vorgelegt und den Beteiligten keinerlei Rückmeldung gegeben. Der MC-Test sollte nicht mit der MC-Übung verwechselt werden.

Der abschließende Nachtest, auch Nachtest 2 genannt, fand 3 oder 4, für ca. 15 % der Probanden erst 7 Tage nach der Übung statt. Den Probanden wurde zunächst mitgeteilt, eine Testung der Bundesstaaten stünde bevor und sie sollten sich diesmal besonders anstrengen, weil sie am Ende Rückmeldung über ihr Gesamtergebnis erhalten würden. Nach der Erhebung aktueller Angst, die im Zusammenhang mit dieser Studie ohne Bedeutung ist, folgte die Testung aller 50 Staaten zunächst mit dem Multiple Choice (im folgenden MC-Test genannt) und anschließend mit Short Answer (im folgenden SA-Test genannt). Beim SA-Test wurden die Items ebenfalls für jeden Probanden nach Zufall dargeboten. Eine korrekte Antwort erforderte exakte Übereinstimmung mit dem Namen des Bundesstaates. Da Testen selbst eine Lernwirkung hervorruft und der MC-Test stets vor dem SA-Test zu bearbeiten war, könnten die Ergebnisse des Short Answer-Tests davon profitieren.

Versuchsplan

Abbildung 2: Formalisierung des Versuchsplans

							unmittelbar	nach 3-7 Tagen
	Gruppe 1:	X_1	X_2	X_3	X_4	O_{MC}	O_{MC}	O_{SA}
R						W		
	Gruppe 2	X_1	X_2	X_3	X_4		O_{MC}	O_{SA}

Das W in Abbildung 2 bedeutet eine Ausbalancierung der Treatmentreihenfolgen und Staatsbereiche über die Probanden. Ein O besteht aus der Testung aller Bundesstaaten, gliedert sich aber in 4 Untertests. Jeder Untertest erfasst diejenigen Staaten, welche den entsprechenden Treatments unterworfen waren. Die Untertests gelten als eigentlich abhängige Variable.

Wie der Versuchsplan verdeutlicht, erlaubt der erste MC-Test bei Gruppe 1 die Prüfung der unmittelbaren, und bei Gruppe 2 der längerfristigen Behaltensleistung der 4

Übungsmethoden via Wiederholungsdesign. Varianzanalytisch kann dieser Versuchsplan gedeutet werden als 2 faktorielles mixed design mit dem within subject factor Übungsmethoden ($X_1..X_4$) und dem between subject factor Testzeitpunkt (unmittelbar, später). Versuchsplantechnisch ist der Faktor Übungsmethoden durch Wiederholungsmessung und der Faktor Testzeitpunkt durch Randomisierung kontrolliert.

Gruppe 1 bearbeitete zusätzlich noch die beiden Behaltenstests MC und SA nach 3 bis 7 Tagen.. Dadurch kann überprüft werden, ob die langfristige Behaltensleistung durch weitere Testung - hier den unmittelbaren Nachtest - verbessert wird. Varianzanalytisch betrachtet: Faktor A: Übungsmethoden ($X_1..X_4$) , Faktor B zusätzliche Testung unmittelbar nach der Übung (ja, nein), abhängige Variable: Posttest nach 3-7 Tagen mit MC und SA.

Hypothesen

Eine zusätzliche Testung führt zu höherem, langfristigen Behalten. Dieser Effekt zeigt sich besonders bei der Übungsmethode "wiederholtes Einprägen", weil diese Bedingungskonstellation dann wenigstens einmal getestet wurde.

Testen mit Feedback führt zu besseren Behaltensleistungen als erneutes Studieren. Dieser Effekt zeigt sich vornehmlich beim langfristigen Behalten, weniger und möglicherweise gar nicht beim unmittelbaren Nachtest, weil bei etlichen Untersuchungen erneutes Studieren zum Teil höheren unmittelbaren Lernerfolg erbrachte als z.B. erneutes Testen. (z.B.: Roediger & Karpicke 2005). Allerdings steht hier ja nicht nur reines Testen, sondern Testen mit Feedback an. Mithin wird erwartet, dass die Übungsvarianten mit Testen und Feedback im Nachtest nach 3 bis 7 Tagen besser abschneiden als erneutes Studieren.

Es liegen keine klaren Erwartungen hinsichtlich der generellen unterschiedlichen Behaltenswirkungen der einzelnen Testübungtmethoden vor. Diesbezüglich hat die Studie explorativen Charakter und dient im wesentlichen der Einschätzung notwendiger und wünschenswerter Anforderungen an Testverfahren bei trivialem Faktenwissen. Die Flashcard-Methoden mit ihrer verstärkten Bearbeitung fehlerhafter Items und dem ausführlicheren Feedback scheinen zwar theoretisch als die effizienteren Übungsverfahren, da sie insgesamt besser gezielte Fehlerkorrekturen ermöglichen. Es ist aber schwer abzuschätzen, in wie weit einfache Wiederholungen mit CSA letztlich den gleichen Effekt bewirken können.

Dennoch werden speziellere Hypothesen formuliert, die jedoch nicht den Kern der Untersuchung ausmachen. So liegt etwa eine testspezifische Wirkung der Übungsmethode auf die Art des Behaltenstests nahe (siehe Jacobs (2006)), etwa in dem Sinne, dass mit MC-Tests eingeübte Staaten besser im MC-Behaltenstest und mit Short-Answer- eingeübte Staaten besser im Short-Answer-Behaltenstest erinnert werden (Kontexteffekt). Ziemlich offensichtlich ist zu erwarten, dass unter der Übungsmethode Covert Short Answer oder MC deutlich mehr Aufgaben pro Zeiteinheit bearbeitet werden als unter Short Answer.

Hinsichtlich der Präferenzen der Studenten für die einzelnen Übungsmethoden wird für die aufwändigeren Methoden eine höhere Wertschätzung prognostiziert und MC mit Feedback die besten Chancen eingeräumt.

Ergebnisse

Vor der Prüfung der Haupthypothesen sei zunächst auf einige Prozessdaten verwiesen, welche nähere Klarheit zu den realisierten Bedingungen erbringen und den Übungsverlauf besser einschätzen lassen.

Wissensniveau zu Beginn der Übung

Eine Analyse des ersten Flashcard-Durchgangs unter der Übungsmethode MC erbrachte eine durchschnittliche Erfolgswahrscheinlichkeit von 49 %. Dieser Wert liefert eine gute Schätzung für das Wissensniveau unmittelbar nach der Lernaneignungsphase und entspricht recht gut der durchschnittlich angestrebten Lernerfolgquote von 50 % zu Beginn der Übungen. Denn so ist zum einen in ca. der Hälfte der Fälle erfolgreicher Retrieval bei den Übungsmethoden mit Testen möglich und es besteht im Mittel noch etliches Lernpotential für alle experimentellen Übungsmethoden.

Bearbeitete Aufgaben für die 4 Übungsvarianten

Bis auf die SA-Übungsvariante waren alle Übungsversionen auf exakt 4 Minuten begrenzt. Nach der ersten Bearbeitung aller Items (12 oder 13) folgte automatisch eine Gesamt- bzw. Teilsicht dieser Items für ca. 20 bzw. 12 Sekunden. Diese Gesamtprozedur wurde sooft wiederholt, bis die Zeitgrenze erreicht war. Je nach Übungsvariante und Lerntempo ließen sich in dieser Zeit unterschiedlich viele Aufgaben bearbeiten. Tabelle 2 stellt die durchschnittliche Anzahl der bearbeiteten Items dar.

Tabelle 2: Anzahl der bearbeiteten Items und Gesamtbearbeitungszeit (N=62)

	Anzahl der bearbeiteten Items		Zeit in Sekunden	
	M	s	M	s
Study only:	56,2	19,4	240	0
Covert short answer:	40,8	16,7	240	0
Multiple Choice:	44,4	14,3	240	0
Short Answer:	37,4	18,1	398	222

Eine durchschnittliche Bearbeitung von 56,2 Items unter Study only bedeutet, dass alle Items nacheinander annähernd 4,5 [56,2/12,5] mal wiederholt zum Einprägen angefordert wurden. Unter der Bedingung Study only wurden hoch signifikant mehr Items bearbeitet als unter allen anderen Übungsversionen. Bei der MC Flashcard-Methode beantworteten die Probanden erwartungsgemäß signifikant mehr Aufgaben als unter klassischer Short Answer-Flashcard-Version, obwohl bei Short Answer mindestens 50% mehr Bearbeitungszeit in Anspruch genommen wurde. Wegen der längeren Übungsphase ist die Short Answer Flashcard-Übung nur bedingt mit den anderen Übungsvarianten vergleichbar und fungiert hier außer Konkurrenz. Hätte man aber die Anzahl der bearbeiteten Items als für alle Methoden konstanter Vergleichsmaßstab gefordert, dann wäre die klassische Short- Answer-Variante benachteiligt.

Reliabilität der abhängigen Variablen:

Während der Übung umfasste jede der 4 Bedingungen 12 oder 13 Bundesstaaten als Items. Alle Bundesstaaten wurden in 4 Subtests eingeteilt, welche diejenigen Items beinhalten, die jeder Proband unter seiner entsprechenden Bedingung einübte. Im spä-

teren Nachtests wurden diese Items sowohl im MC wie im SA-Format getestet. Wie die Korrelationen in Tabelle 3 aufzeigen, messen beide Testformen in hohem Maße Vergleichbares. Die Höhe der Korrelationen lässt keinen Zweifel an der Zuverlässigkeit jedes Messinstruments aufkommen.

Tabelle 3: Korrelationen zwischen MC- und Short Answer Nachtests für die einzelnen experimentellen Übungsbedingungen (N=54)

Multiple Choice	**.86**
Study only	**.85**
Covert short answer	**.81**
Short Answer	**.76**

.86 bedeutet. Die Subtests MC und SA für die Übungsbedingung Multiple Choice korrelieren .86 miteinander.

26 Probanden absolvierten den MC-Test unmittelbar, sowie mindestens 3 Tage später. Die Retestkorrelationen der 4 MC-Subtests schwanken zwischen r=.70 und r=.76, sind alle hochsignifikant und bestätigen die Zuverlässigkeit des MC-Tests.

Führt das Testen nach einer Übung zu einem verbesserten langfristigen Behalten ?

Im Versuchsplan war festgelegt worden, dass 2 Probanden exakt dieselben Bedingungen (z.B. Reihenfolge der Übungsmethoden) erhielten und der Zufall dann entscheidet, welcher der beiden Probanden den unmittelbaren Nachtest durchführt.

Mit Hilfe des erfragten Abiturnotendurchschnitts sollte unter anderem der Erfolg der Randomisierung überprüft werden. Entgegen jeder Erwartung fielen die Abiturdurchschnittswerte der Gruppe 1 (unmittelbare Testung) signifikant besser aus als die der Gruppe 2 (keine unmittelbare Testung) ($t(52) = -2.44$, $p<.05$, Effektstärke=d=.65) , was die interne Validität dieses Vergleichs insofern etwas gefährdet als die Abiturdurchschnittsnote hochsignifikant mit r = -.37 (N=54) bzw. r = -.40 (N=53) mit den beiden abhängigen Variablen, MC und SA-Posttest, korreliert. Die Randomisierung schließt derartige Zuteilungen leider nicht aus. [Die unerwartete Aufteilung der Probanden durch Randomisierung ist mir allerdings bei dem vorhergehenden Experiment mit einfacher Randomisierung ebenfalls begegnet, was nun sicherlich signifikant ist und mich als den Pechvogel der Randomisierung bestätigt.]

Die unmittelbar getesteten Probanden konnten ihre Testbearbeitungszeit selbst bestimmen und benötigten im Durchschnitt ca. 5 Minuten (genaue Werte in Sekunden: M=306, s=78, N=31). Für die nicht getesteten Probanden war die Übung hingegen schon beendet. Betrachtet man die Nachtestung selbst als einen Teil der Übung, so lässt sich in Tabelle 4 die gesamte reine Übungszeit für beide Gruppen wie folgt in Minuten aufschlüsseln:

Tabelle 4: Übungszeiten für Probanden mit und ohne unmittelbaren Nachtest

	Lernan eignung	Experiment Übung	Test	insgesamt
Test	10,5	18,6	5	34,1
kein Test	10,5	18,6	0	29,1

Die nicht getestete Gruppe benötigte somit im Vergleich zur getesteten Gruppe nur ca. 85% der Übungszeit. Zunächst soll überprüft werden, ob eine Testung (hier der Nachtest 1) im Anschluss an eine Übung das Behalten im Nachtest einige Tage später (im folgenden auch Nachtest 2 genannt) stärkt. Entsprechend dem Versuchsplan werden

die Ergebnisse im Nachtest 2 für die zuvor getestete Gruppe mit denen der zuvor nicht getesteten Gruppe verglichen.

Ein zusätzlicher Test sollte bei allen Übungsmethoden die Behaltensleistung fördern, besonders jedoch bei Study only, weil bei dieser Übungsmethode niemals ein echter Retrieval verlangt wurde, der ein Abruftraining fördern würde.

Die zweifaktorielle Varianzanalyse mit den 4 Übungsmethoden als within subject factor und der vorherigen Testung (ja,nein) als between subject factor ergab für den Hauptfaktor vorheriges Testen folgende Ergebnisse im Hinblick auf beide Kriteriums-variablen im Nachtest 2.

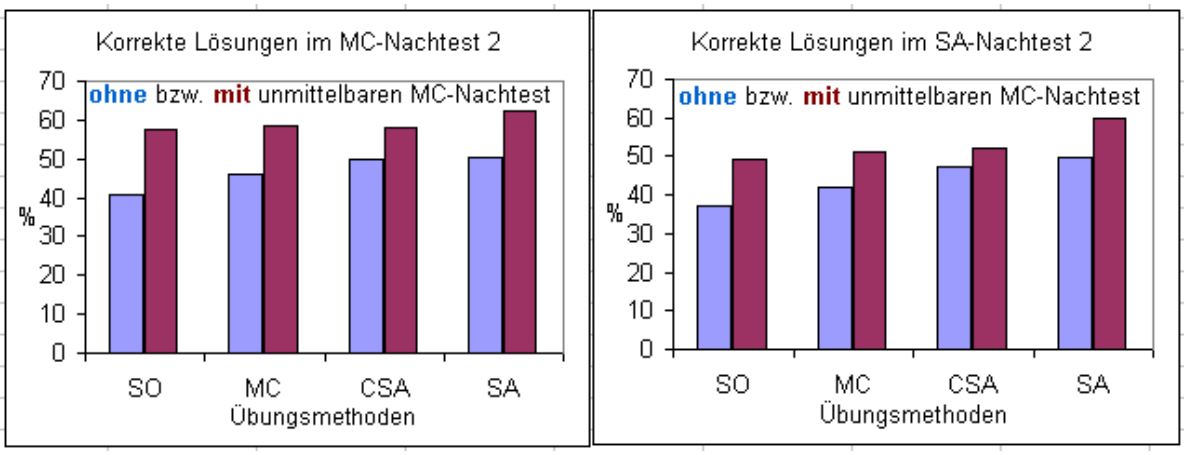
MC: $F(1,53) = 5,87$; $p=.019$; partielles EtaQuadrat= .1:
SA: $F(1,52) = 2.37$; $p=.129$ partielles EtaQuadrat= .04:

Tabelle 5 und Abbildung 3 zeigen die Ergebnisse im MC- und SA-Test getrennt für alle untersuchten Übungsmethoden und unterteilt in Personen, die den unmittelbaren MC-Test absolviert haben (ja) oder nicht (nein)

Tabelle 5: Prozentsatz korrekter Lösungen im Nachtest 2 (ca. 4 Tage nach der Übung) für alle untersuchten Probanden.
N pro Gruppe 26 - 28; N Gesamt 54-56

		MC-Test			SA-Test	
	Test	M	*s*		M	*s*
Study only (SO)	ja	**57,4**	*26,0*		**49.1**	*29,5*
	nein	**40,7**	*21,4*		**37,1**	*22,1*
	Gesamt	**48,9**	*25,0*		**43,1**	*26,5*
Multiple Choice (MC)	ja	**58,4**	*25,6*		**50,9**	*28,6*
	nein	**46,0**	*23,9*		**42,2**	*24,7*
	Gesamt	**52,1**	*25,3*		**46,6**	*26,9*
Covert short Answer (CSA)	ja	**58,2**	*28,9*		**52,0**	*29,0*
	nein	**49,7**	*22,9*		**47,1**	*17,9*
	Gesamt	**53,9**	*26,1*		**49,2**	*24,7*
Short Answer (SA)	ja	**62,5**	*22,7*		**59,9**	*24,6*
	nein	**50,3**	*20,2*		**49,9**	*25,3*
	Gesamt	**56,3**	*22,1*		**54,9**	*25,2*

Abbildung 3: Ergebnisse für alle experimentellen Bedingungen in beiden Nachtests ca. 4 Tage nach der Übung



Konsistent über alle Übungsmethoden hinweg erzielten die zuvor getesteten Probanden im Nachtest 2 numerisch höhere Behaltenswerte. Wie Tabelle 6 belegt, beträgt der Behaltensvorteil im Durchschnitt aller Übungsmethoden für den MC-Test 12% (Effektstärke d=.65) und für den SA-Test 9% (Effektstärke d=.42).

Tabelle 6: Prozentsatz korrekter Lösungen für alle Übungsmethoden im Nachtest, 3 bis 7 Tage nach der Übung

Nachtest-format	unmittelbar nach der Übung	N	M	s	t	d
Multiple Choice	MC-Test	27	59,3	22,2	2,42	.65
	Kein MC-Test	28	46,9	15,4		
Short Anwer	MC-Test	27	53,2	24,9	1,55	.42
	Kein MC_Test	27	44,2	16,8		

Allerdings lässt sich der Testeffekt im Mittel aller Übungsmethoden wegen der überaus hohen Personenstreuung nur beim MC-Test statistisch eindeutig sichern, was ja bereits aus dem Ergebnis der VA hervorging. Als unmittelbarer Nachtest wurde auch lediglich der MC-Test vorgegeben und vielleicht zeigt sich deshalb der Testeffekt hier besonders klar. Außerdem könnte der MC-Test eine nivellierende Wirkung auf den SA-Test ausüben, weil der MC-Test selbst eine Übungsgelegenheit bietet und stets vor dem SA-Test bearbeitet wurde.

Wie man der Abbildung 3 entnimmt, erscheint der Unterschied bei SO etwas größer als bei manchen Testübungsmethoden. Die Interaktion zwischen Übungsmethoden und Testen erreicht jedoch in keinem Falle auch nur ansatzweise die Signifikanzhürde. Somit lässt sich auf der Basis der Varianzanalyse nicht behaupten, die Wirkung der vorherigen Testung würde sich bei den einzelnen Übungsmethoden unterschiedlich auswirken. Dennoch bringt eine weniger strenge, differenzierte Analyse weiteren Auf-

schluss. Tabelle 7 stellt das Ergebnis des t-Test der entsprechenden Mittelwertsunterschiede dar. Dabei wurde eine Alphafehlerkumulierung in Kauf genommen.

Tabelle 7: Mittelwertsvergleich der zuvor mit MC getesteten Gruppen und der zuvor nicht getesteten Gruppen 3 bis 7 Tage nach der Übung im SA und MC-Nachtest, nach Übungsmethoden differenziert. (df = 52-53)

		unmittelbar nach der Übung MC-getestet			
	Nach-test2	ja	nein	t	p
Study only	**MC**	**57,4**	**40,7**	**2,60**	**.006**
	SA	**49,1**	**37,1**	**1,69**	**.049**
Multiple Choice	MC	58,4	46,0	1,86	.034
	SA	50,9	42,2	1,21	.12
Covert Short Answer:	MC	58,2	49,7	1,22	.12
	SA	52,0	47,1	0,75	.23
Short Answer:	MC	62,4	50,1	2,11	.02
	SA	59,9	49,9	1,47	.07

Immerhin lässt sich der Unterschied zwischen den zuvor getesteten und nicht getesteten Probanden für die Methode SO gesondert sowohl für den langfristigen MC-Nachtest wie für den Short Answer-Nachtest 3 bis 7 Tage nach der Übung statistisch belegen. Bei der Übung CSA hingegen hatte eine unmittelbare Testung nach der Übung bei keiner AV einen signifikanten Effekt auf das langfristige Behalten. Die MC- sowie die Short Answer Übungsmethoden zeigen lediglich einen signifikanten Vorteil der getesteten Gruppe im MC-Test.

Insgesamt bestätigen die Befunde die das Behalten fördernde Wirkung eines Tests im Anschluss an eine Übung. Die Testwirkung zeigt sich eindeutig bei der Übungsmethode "erneutes Studieren", wo sie auch theoretisch am ehesten vermutet worden war. Die Durchführung eines reinen Tests ohne Feedback verbesserte demnach das Behalten und kann hier als Teil einer effektiven Übung betrachtet werden. Begünstigt wurde dieses Ergebnis vermutlich durch die relativ hohen Erfolgsquoten im unmittelbaren Nachtest von ca. 70 Prozent (siehe weiter unten). Denn dadurch vollzogen die Probanden in den meisten Fällen einen positiven Abrufprozess bei der gestellten Frage. Es ist schwer einzuschätzen, ob zusätzliches KCR-Feedback im gegebenen Fall noch deutlich bessere Behaltenswerte nach sich gezogen hätte, weil das Korrekturpotenzial mit knapp unter 30 Prozent relativ begrenzt erscheint.

Fördert Testen mit Feedback den Lernerfolg mehr als erneutes Studieren?

Fördert Testen mit Feedback den unmittelbaren Lernerfolg mehr als erneutes Studieren?

Nach der Übung sowie einer sich anschließenden ca. 2 bis 4 Minuten dauernden Befragung bearbeitete ein Teil der Probanden lediglich den MC-Behaltenstest, der hier als unmittelbarer Nachtest bezeichnet wird. Wie bei der Hypothesenformulierung erwähnt, waren unmittelbar nach der Übung nicht in jedem Falle bedeutsame Unterschiede zwi-

schen den Übungsmethoden erwartet worden, da bei jeder Methode intensiv geübt wurde, was zumindest kurzfristiges Behalten ermöglichen sollte.

Tabelle 8: Prozentsatz der korrekten Lösungen im unmittelbaren MC-Nachtest (N=31)

	M	s
Study only:	69.7	24.2
Multiple Choice:	68.6	26.1
Covert short answer:	72.2	22.1
Short Answer:	71.2	18,1

Die einfaktorielle Varianzanalyse mit den Übungsmethoden als Wiederholungsfaktor erbrachte keinen signifikanten Haupteffekt ($F(3,28)=0.316$ $p=.81$.) sowie keinerlei Unterschiede zwischen allen möglichen Vergleichen. Wie auch die Tabelle 8 überzeugend aufzeigt, unterscheiden sich die verschiedenen Übungsmethoden somit nicht im Hinblick auf die Förderung des unmittelbaren Behaltens. Das Ergebnis fällt nicht sehr überraschend aus. Der durch die experimentellen Übungsvarianten bewirkte Lerngewinn nach der Lernaneignungsphase kann auf etwa 20 % eingeschätzt werden.

Fördert Testen mit Feedback das langfristige Behalten mehr als erneutes Studieren ?

Eine unmittelbare Testung im Anschluss an die Übung führt zu Veränderungen, welche z.B. hier nachweislich Einfluss auf das langfristige Behalten ausübte und die Erfassung der ursprünglichen Übungswirkung bei allen Übungsvarianten verändert. Nur bei den Probanden, welche den unmittelbaren Nachtest nicht absolvierten, kann etwa die Auswirkung der Study Only Bedingung in Reinkultur für längerfristiges Behalten überprüft werden, da unter dieser Bedingungskonstellation immer nur Informationen präsentiert wurden und der nach ca. 4 Tagen anberaumte Nachtest 2 die erste Testung für diese Items darstellt. **Deshalb beschränkt sich nachfolgende Analyse auf die Probanden, welche nach der Übungsphase keinen unmittelbaren Nachtest bearbeiteten.**

Die Hypothese lautet eindeutig, Testen mit Feedback führe zu höherem langfristigen Behalten als erneutes Studieren. Die beste Möglichkeit, diese Frage zu überprüfen, liefert der Vergleich zwischen Study Only und Covert Short Answer. Denn beide Übungsmethoden unterscheiden sich lediglich dadurch, dass unter CSA zunächst ein Fragezeichen im Staatsgebiet erschien, welches zur Erinnerung an den Staatsnamen aufforderte. Unter der SO-Bedingung sahen die Probanden den Namen des Staates direkt im entsprechenden Territorium, unter CSA-Bedingung folgte die Präsentation des Namens erst im Anschluss an die gedachte Antwort als Feedback KCR. Die erwartete Überlegenheit der Testübungsgruppen gilt natürlich auch für die Vergleiche SO mit MC und SA, da letztere Testmethoden ja theoretisch als verbesserte Testübungsvarianten konzipiert wurden. Tabelle 9 verdeutlicht neben den deskriptiven Ergebnissen, dass die Varianzanalyse mit Messwiederholung lediglich beim Short Answer-Nachtest 2 einen signifikanten Haupteffekt für die Übungsmethoden verzeichnete. Tabelle 10 stellt die Ergebnisse der t-tests bzgl. der theoretisch entscheidenden Hypothesenprüfungen unter Inkaufnahme einer Kumulierung des Alphafehles dar.

Tabelle 9: Mittelwerte und Streuungen der Übungsmethoden 3 bis 7 Tage nach der Übung nur für Probanden ohne unmittelbaren Nachtest

	Multipe Choice		Short Answer	
	M	*s*	M	S
Study only	40.7	*21.4*	37.1	*22.1*
Multiple Choice	46.0	*24.4*	42.2	*24.7*
Covert short Answer	49.7	*22.9*	47.1	*17.9*
Short Answer	50.3	*20.2*	49.9	*25.3*
Haupteffekt Übungsmethoden	F(3.81)=1.62 p =0.19		F(3.78)=2.79 p=0.046	

Tabelle 10: t-Test sowie Effektstärke d zwischen SO und allen Testübungsmethoden

	Nachtest, 3 bis 4 Tage nach der Übung							
	Multiple Choice				Short Answer			
	t	df	p	d	t	df	p	d
Multiple Choice vs. Study only	1.07	27	.14	.23	1.27	26	.11	.22
Covert Short Answer vs. Study only	**1.96**	**27**	**.03**	**.41**	**2.24**	**26**	**.017**	**.50**
Short Answer vs. Study only	2.50	27	.01	.46	2.96	26	.004	.54

Wie aus Tabelle 10 hervorgeht, lässt sich der langfristige Behaltensvorteil von CSA gegenüber SO für beide Kriteriumstests signifikant bestätigen. Dieser Nachweis gelingt allerdings erwartungswidrig nicht, wenn man die SO-Übungsmethode mit der MC-Übungsmethode vergleicht. Der offensichtliche, stets hochsignifikante Vorteil von SA gegenüber SO ist hier von geringerer Bedeutung, da die Lernzeiten nicht vergleichbar sind, wenngleich sich so der Nutzen intensiven SA-Testens offenbar auszahlt. Unter SA wird zwar hochsignifikant länger geübt als bei SO, aber es werden auch hochsignifikant weniger Items bearbeitet (siehe Tabelle 3). Insgesamt kann die Untersuchung einige empirische Belege dafür anführen, vermehrtes Testen mit Feedback sei dem wiederholten Studieren hinsichtlich eines längerfristigen Behaltens überlegen.

Der Vergleich zwischen SO and CSA hat hohe Ähnlichkeit mit dem Vorgehen von Carrier und Pashler (1992). Denn dort wurde SO ebenfalls mit einer vereinfachten SA-Variante verglichen. Die Anzahl der Items und die Darbietungszeit für jedes Item war unter beiden Bedingungen gleich gehalten. Hier jedoch konnten die Probanden bei exakt gleicher Studierzeit für beide Bedingungen selbst entscheiden, wie viele Items sie in welcher Geschwindigkeit bearbeiten. CSA hat sich hier gegenüber SO beim langfristigen Behalten als überlegen erwiesen, obwohl die Probanden unter SO mindestens einen Übungsdurchgang mehr absolviert hatten. (siehe Tabelle 2). Es kommt offenbar nicht nur auf die Informationsmenge, sondern auch auf die Art bzw. Güte der Informationsverarbeitung an.

Welche Testübungsmethode erbringt den höchsten langfristigen Behaltensgewinn?

Tabelle 5 und Abbildung 3 zeigen insgesamt nur geringe Unterschiede zwischen den verschiedenen Testübungsvarianten im Hinblick auf das langfristige Behalten auf. **Insbesondere die sehr aufwändigen Übungsvarianten MC und SA schneiden keineswegs besser ab als CSA.** Die sich gelegentlich andeutende Überlegenheit von SA im Vergleich zu MC (z.B. beim SA-Nachtest) ist wegen des höheren Zeitbedarf bei SA nicht schlüssig interpretierbar, auch wenn unter MC mehr Items bearbeitet wurden. Desgleichen lassen sich keine klaren Kontexteffekte ausmachen, die darauf hindeuteten, die speziellen Übungsvarianten MC und SA würden sich als vorteilhafter bei den entsprechenden Testvarianten MC bzw. SA erweisen. Unter Abwägung von Aufwand und Ertrag hat sich CSA im gegebenen Fall als eine hinreichend gute Testvariante erwiesen.

Subjektive Einschätzung der Übungsmethoden

Nach der Übung wurden die Probanden aufgefordert, die einzelnen Übungsmethoden mit Noten zu bewerten. Anschließend hatten sie Aufgabe, die Übungsmethoden entsprechend Ihrer Präferenz in eine abfallende Rangfolge zu bringen, wobei die Ränge wie folgt verankert waren. 1 = beste Methode; 2 = zweitbeste Methode; 3 = drittbeste Methode; 4 = am wenigsten geeignete Methode). Beide Einschätzungsvariablen messen Vergleichbares in akzeptabler Zuverlässigkeit. Denn die Korrelationen beider Maße bewegen sich für die einzelnen Übungsmethoden in einem Bereich zwischen .63 und .90. Tabelle 11 listet Mittelwerte und Standardabweichungen für beide Bewertungsvariablen auf und fasst die statistischen Unterschiede zusammen.

Tabelle 11: Benotung und Rangfolge der Übungsmethoden durch die Probanden (N=66)

	Note		Rang	
	M	s	M	s
Multiple Choice	**1,9**	0,8	**2,0**	0,9
Short Answer	**2,3**	1,2	**2,5**	1,2
Covert short answer	**2,4**	1,0	**2,5**	1,1
Study only	**2,8**	1,2	**3,0**	1,2
	MC<(SA=CSA)<SO		MC<(SA=CSA)<SO	

< signifikant geringer .05% zweiseitig; = nicht signifikant; jeweils nach t-Test für abhängige Stichproben

Die Multiple Choice Übung (mit Feedback und Flashcard) wird eindeutig am besten und die Study-Only-Methode ziemlich klar am schwächsten eingeschätzt, während beide Short Answer Versionen in der Mitte liegen. Der Unterschied in der Bewertung zwischen MC-Testung und Study-Only fällt mit einer Effektstärke d um ca. 1 recht deutlich aus.

Leistungseinschätzungen unmittelbar nach der Übung

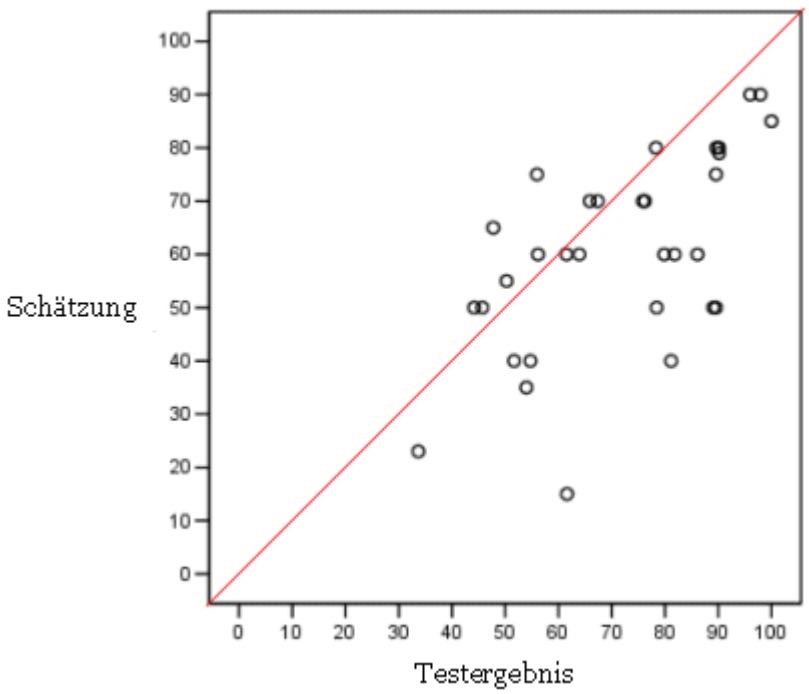
Die unmittelbar nach der Übung folgende Befragung beinhaltete auch eine subjektive Einschätzung des Behaltens in Form der Frage: "Stellen Sie sich vor, Sie müssten die Namen aller amerikanischen Bundesstaaten auf einer Landkarte zutreffend positionieren. Schätzen Sie bitte ein. Wie viel Prozent würden Sie aktuell korrekt zuordnen ? --- % richtig". Die einzuschätzende Behaltensleistung entsprach von den Anforderungen exakt der sich unmittelbar anschließenden, tatsächlichen Testung. Tabelle 12 stellt die wesentlichen Befunde zusammen.

Tabelle 12: Objektive Leistungsdaten und subjektive Leistungseinschätzung der Probanden in Prozent der Lehrzielerreichung

	MC Test objektiv	Schätzung subjektiv
M	71.4	60.5
s	18.5	18.3
	t (31) = 3.8 , p<001	
Korrelation	=.63	

Objektive Behaltensleistung und subjektive Schätzung korrelieren zwar erwartungsgemäß hochsignifikant, unterscheiden sich aber deutlich im Niveau. Die signifikante Unterschätzung der objektiven Behaltensleistung deutet auf einen "underconfidence with practice effect" hin (Koriat, Sheffer.& Ma'ayan (2002)}. Allerdings wird aus der Abbildung 4 erkennbar, dass - sieht man von einem krassen Ausreißer (60/15) ab- hauptsächlich die sehr Leistungsfähigen für die Unterschätzung verantwortlich sind. Deren subjektiver Spielraum nach oben ist natürlich auch recht begrenzt. Wie schon in etlichen Untersuchungen (z.B. Jacobs 2003, Clayson 2005) gefunden, steigt die Unterschätzung mit wachsender objektiver Leistung, was sich auch in der negativen Korrelation zwischen der Differenz (Schätzung - Ergebnis) und objektiven Testergebnis von r = - .43 niederschlägt, die nach Elimination des Ausreißers bereits r=-.51 (N=31) beträgt.

Abbildung 4: Zusammenhang zwischen Testergebnis und Schätzung



Diskussion

Die Studie erbrachte weitere Belege für die Lernwirksamkeit reinen Testens sowie die Nützlichkeit des Testens mit Feedback für das längerfristige Behalten. Zu den wichtigsten Ergebnissen zählen:

1. Ein zusätzlicher Test (ohne Feedback) am Ende einer Übung verbessert das langfristige Behalten.
2. Testen mit Feedback stärkt das Behalten wirksamer als wiederholte Informationsaufnahme.
3. Ein einfaches Test- und Feedbackverfahren reicht als Testübungsmethode aus.
4. Subjektiv einschätzte Lernwirksamkeit und objektiver Lernzuwachs stimmen nur teilweise überein.

Es sei ausdrücklich darauf hingewiesen, dass die Erfolgswahrscheinlichkeit beim unmittelbaren Nachtest mit 71 % relativ hoch ausfiel. Das Testen (ohne Feedback) führte somit zu einem hohen Anteil erfolgreicher Abrufprozesse, was meiner Meinung nach eine notwendige Voraussetzung für den verantwortungsvollen Einsatz reinen Testens als Übungsmethode ist. Der Testeffekt zeigte sich besonders klar bei der Study Only Methode, die in der Übung ausschließlich wiederholte Informationsaufnahme und keinerlei Testung beinhaltete. Weil reines Testen unter günstigen Bedingungen mehr Lerngewinn bewirkt als gar nichts, lohnt es sich zwar, zu testen. Man kann aber mit Recht daran zweifeln, ob eine No-treatment-Kontrollgruppe als hinreichend ernsthafte Vergleichsbasis betrachtet werden darf, wenn andere nahe liegende Maßnahmen einen ähnlichen oder gar höheren Lerngewinn versprechen könnten. Überzeugender wäre zweifellos ein Nachweis, reines Testen sei wichtiger als erneutes Einprägen. Ich nehme an, dass dies nur unter ganz besonderen Bedingungen, z.B. sehr hohen Erfolgswahrscheinlichkeiten, der Fall sein wird. Die Ergebnisse der Experimente 3 und 4 von Cull (2000) deuten darauf hin, wiederholtes reines Testen sei dann dem wiederholten Studieren überlegen, wenn sich einer sehr guten Lernaneignung unmittelbar eine massierte Übung anschließt und die Behaltensleistung nach einem längeren Retentionsintervall erfasst wird. Die Befunde von Roediger und Karpicke (2005) zum Free Recall Testen deuten in die gleiche Richtung.

Der eigentlich interessante Vergleich lautet jedoch: "Erneutes Studieren gegen Testen mit Feedback". Das zweite wichtige Ergebnis dieser Studie erscheint in diesem Zusammenhang deshalb aussagekräftiger. Die Übungszeiten waren hier absolut vergleichbar, Testen mit Feedback aber dennoch wiederholtem Studieren überlegen. Auch wenn nicht alle Übungsvarianten mit Tests und Rückmeldung bei jeder abhängigen Variable einen signifikanten längerfristigen Behaltensvorteil gegenüber erneutem Studieren erbrachten, liegen alle möglichen Ergebnisse in erwarteter Richtung.

Die erwarteten Unterschiede zeigten sich nur beim langfristigen Behalten. Dieses Ergebnis kommt zwar nicht ganz überraschend , da etliche Studien nur für langfristiges Behalten Vorteile des Testens (mit Feedback) ergaben, steht aber im Widerspruch zu dem Experiment von Carrier und Pashler (1992) sowie Cull (2000, Experiment 2), die auch Vorteile des Testens mit Feedback für unmittelbare Behaltensleistungen fanden.

Covert Short Answer mit KCR-Feedback hat sich als effiziente Testübungsmethode erwiesen. Wie schon De Klerk & De Klerk (1978) nachgewiesen haben, kann es genügen, sich die korrekte Antwort im Geiste vorzustellen, wenn anschließend mindestens KCR folgt. Die zusätzlichen, theoretisch durchaus begründbaren Features der aufwändigen MC- und SA- Übungsmethoden mit Flashcard und partiellem Response Contin-

gent Feedback haben keine zusätzlichen Verbesserungen erbracht. Unter der MC-Übungsvariante haben die Probanden 21% aller Items verwechselt, unter der Short-Answer-Variante lediglich 10%, weil sie den falschen Staat exakt eingeben mussten. In allen diesen Fällen wurde Response Contingent Feedback wirksam, indem zur korrekten Antwort (KCR) zusätzlich noch der falsch angegebene Staat zutreffend rückgemeldet wurde. Das Prinzip von Flashcard führte sicherlich zu einer höheren Anzahl von falsch bearbeiteten Items während der Übung, die durch das KCR-Feedback ein erneutes Enkodieren anregen konnten. Aber weder die erhöhte Anzahl der schwierigen Itemvorgaben mit Rückmeldung, noch das relativ seltene RCF reichten offenbar aus, um sich als klare Verbesserung gegenüber einfachem KCR im Gesamtergebnis auszuweisen. Möglicherweise setzen sich diese Sonderfaktoren erst bei größeren Itemmengen und recht beschränkter Übungszeit durch, weil dann ein höheres relatives Korrekturpotential verfügbar ist. Hier aber konnte der Lerner unter CSA die beherrschten Items schnell wegklicken, sich mehrmals intensiv den schwierigen Items zuwenden, die notwendige Rückmeldung dazu einfordern und die Zeit zum verbesserten Enkodieren selbst bestimmen. Bei frei wählbarer Übungszeit scheinen CSA-analoge Aufgabentypen durchaus als ernsthafte Übungsalternativen zu formalen Testversionen zu fungieren. (siehe etwa Concept Identifikation Exercise). Mit zunehmender Anzahl von Übungsdurchgängen wird es immer schwerer, durch zusätzliche Feedbackverbesserungen einfaches KCR an Wirksamkeit zu übertreffen.

Das Experiment macht keine Aussage darüber, ob eine Computer gestützte Fassung letztlich mehr Behaltenserfolg verspricht als herkömmliche Methoden des Erlernens, etwa ganz konventionell mit Hilfe eines Atlas. Denn auch bei der Methode Study Only organisierte der Computer die Itemauswahl und bot gewisse Interaktionsmöglichkeiten an. Eine für die Zukunft geplante Fragestellung bezieht sich auf die fremd oder selbst bestimmte Auswahl der zu bearbeitenden Aufgaben. Wird das Einprägen wirkungsvoller gefördert, wenn der Computer in einer gewissen Systematik Aufgaben vorgibt oder weiß der Lerner selbst es gar besser, welche Aufgaben er in welcher Intensität bearbeiten soll?

Die subjektiven Einschätzungen zur Wirksamkeit der Übungsmethoden entsprechen nur teilweise den objektiven Behaltenswerten, ein Befund, der erneut auf die Wichtigkeit aufmerksam macht, sich bei der Bewertung von Software nicht zu sehr auf subjektive Einschätzungen zu verlassen. Zutreffend bewerteten die Probanden Study Only als die schwächste Lernmethode. Die signifikant bessere Einschätzung von MC gegenüber den beiden Short-Answer-Versionen entspricht aber nicht den objektiven Relationen. Das Anklicken beim MC-Aufgabentyp lässt sich recht bequem bewerkstelligen und die diversen Rückmeldungen gewähren einen komfortablen Feedbackservice, was die Lerner offenbar schätzen. Wenn man dann eher freiwillig Faktenwissen einübt, so hätte die spezielle MC-Flashcard-Methode immerhin motivationale Vorteile.

Die hier durch die unterschiedlichen Übungsmethoden bewirkten Behaltensunterschiede erklären in der Regel weniger als 10% der Varianz. Erheblich mehr Varianz, teilweise mehr als 50% geht auf die Personen zurück. Manche können sich die Bundesstaaten gut einprägen, andere nicht, gleichgültig mit welcher Methode. Der Abiturnotendurchschnitt "erklärt" - man kann selbstverständlich nicht sagen bewirkt - deutlich mehr Behaltensvarianz als die unterschiedlichen Übungsmethoden. Er lässt sich allerdings nicht experimentell variieren.

Schon unmittelbar nach der Übung und vor dem unmittelbaren Nachtest setzte das Vergessen ein. Vom unmittelbaren Nachtest an sank die Erfolgsquote innerhalb von ca. 4 Tagen im Durchschnitt um 10 bis 15 Prozent bei den Probanden, welche am Nachtest teilnahmen. Bei den nicht unmittelbar getesteten Probanden ist die Vergessensquote im

entsprechenden Zeitraum bereits auf mehr als 20% zu veranschlagen. Alles, was nicht mindestens gelegentlich wiederholt wird, fällt langfristig dem Vergessen anheim.

Zu meiner eigenen Überraschung schätzen die Probanden das auf unterstem Lehrzielniveau angesiedelte Übungsprogramm "Einprägen der Bundesstaaten der USA" eher als interessant, motivierend und anregend ein. Sie waren von der Behaltenswirksamkeit des Übungsprogramms überzeugt. Zum Teil führe ich diese relativ gute Bewertung auf den erlebten Methodenwechsel innerhalb der Gesamtübung zurück.

Verbesserungsmöglichkeiten des Einübens einfachen Faktenwissens sehe ich vornehmlich darin

- die Übungen auf mehrere Zeitpunkte zu verteilen,
- Übungsmaßnahmen vielfältig zu variieren,
- die Fakten selbst mit mehr Sinn zu verbinden.

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, Vol 52(4), 566-577.

Bangert-Drowns, R.L., Kulik, C., Kulik, J.A., & Morgan, M.T. (1991). The instructional effect of feedback in test-like events. Review of Educational Research,61, 213-238.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. Memory & Cognition, 20, 632-642.

Carpenter, S. K. & DeLosh, E.L. (2005). Application of the testing and spacing effects to name learning. Applied Cognitive Psycholy (in press)
Published Online: 14 Mar 2005
URL: http://www3.interscience.wiley.com/cgi-bin/abstract/110430132/ABSTRACT [4.4.2005]

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T. & Rohrer, D. (in press). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin.

Clayson, D. E. (2005). Performance Overconfidence: Metacognitive Effects or MisplacedStudent Expectations? Journal of Marketing Education, Vol. 27, No. 2, 122-129

Cull, W. L. (2000). Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall. Appl. Cognit. Psychol. 14: 215-235

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]

De Klerk, L. F. W. & De Klerk, L. (1978). The effect of knowledge of correct results per item on verbal learning and retention. Instructional Science. Vol 7, Nr. 4, 347 – 358.

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

Dempster, F. N., Dunbar, M. E., Corkill, A.J. : University of Nevada, Las Vegas (2001). Explorations in the Effective Use of Multiple Learning Opportunities:
Study Versus Test Conditions
http://edtech.connect.msu.edu/Searchaera2002/viewproposaltext.asp?propID=5384 [13.12.2005]

Duchastel, P. C. and Nungester, R. J. (1982). Testing effects measured with alternative test forms. Journal of Educational Research, vol. 75, no. 5, pp. 309 -313

Glover, J. A. (1989). The testing phenomenon: Not gone but nearly forgotten. Journal of Educational Psychology, 81, 392-399.

Hamaker, Ch. (1986). The Effects of Adjunct Questions on Prose Learning. Review of Educational Research, Vol.56, No 2, Pp 212-242.

Haynie , W. J. (1994). Effects of Multiple-Choice and Short-Answer Tests on Delayed Retention Learning Journal of Technology Education Volume 6, Number 1
http://scholar.lib.vt.edu/ejournals/JTE/v6n1/haynie.jte-v6n1.html
http://scholar.lib.vt.edu/ejournals/JTE/v6n1/pdf/haynie.pdf

Jacobs, B. (1998-2005) Entwurf für eine Studie zum Nachweis der Wirksamkeit des KCR-Feedbacks im Anschluss an korrekte Antworten.
http://www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/kcr-entwurf.htm

Jacobs, B. (2001). Die Wirkung von Lösungsbeispielen, Aufgaben und Feedback auf das Lösen von Kombinatorikproblemen.
URN: urn:nbn:de:bsz:291-psydok-3105
URL: http://psydok.sulb.uni-saarland.de/volltexte/2004/310/

Jacobs, B. (2002). Aufgaben stellen und Feedback geben.
URN: urn:nbn:de:bsz:291-psydok-4387
URL: http://psydok.sulb.uni-saarland.de/volltexte/2004/438/

Jacobs, B. (2003). Feedback mit oder ohne eigene Aufgabenbearbeitung ?
http://www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/direktesfeedback.htm

Jacobs, B. (2003b). Leistungsabhängige Selbsteinschätzungen unmittelbar vor einer Klausur.
URN: urn:nbn:de:bsz:291-psydok-399
URL: http://psydok.sulb.uni-saarland.de/volltexte/2003/39/

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. (2005a). Fragen stellen und Rückmeldung geben beim Paarassoziationslernen
http://www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/paarassoziation_kcr.htm

Jacobs, B. (2005b). Die Wirkung sehr offener Fragen auf das Behalten eines Lehrtextes
http://www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/free_recall_testen.htm

Jacobs, B. (2006). Die Auswirkungen von Short-Answer- und Multiple Choice-Übungsaufgaben auf das Lernergebnis.
http://www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/sa_gegen_mc.html

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2005, May). Testing enhances memory retention, but which test format is better? [*Poster presented at the 17th American Psychological Society Annual Convention, Los Angeles, CA.*]

Kuo, T-M., & Hirshman, E. (1996). Investigations of the testing effect. American Journal of Psychology, 109, 451-464.

Koriat, A. , Sheffer, L. & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. Journal of Experimental Psychology: General, Vol 131(2), pp. 147-162.

Pashler, H., Cepeda, N. J., Wixted, J. T. & Rohrer, D. (2005). When Does Feedback Facilitate Learning of Words? Journal of Experimental Psychology: Learning, Memory, and Cognition 2005, Vol. 31, No. 1, 3 8

Pashler, H.; Zarow, G.; Triplett, B. (2003). Is Temporal Spacing of Tests Helpful Even When It Inflates Error Rates? Journal of Experimental Psychology / Learning, Memory & Cognition, Nov2003, Vol. 29 Issue 6, p1051-1057, 7p;

Morris, P. E., Fritz, C. O. ,Jackson, L., Nichol, E. & Roberts, E. (2005). Strategies for

Learning Proper Names: Expanding Retrieval Practice, Meaning and Imagery Applied Cognitive Psychology 19: 779 798 .

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. Journal of Educational Psychology, 74(1), 18-22.

Rawson, K. A. & Kintsch, W. (2005). Rereading effects depend on time of test. Journal of Educational Psychology, 97, 70-80.

Roediger, H. L. & Karpicke, J. D.(2005). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychological Science (in press)
http://psych.wustl.edu/memory/Roddy%20article%20PDF's/Roediger%20Karpicke%20Preprint%20PsychSci.pdf [3.5.2005]]

Richland, L. E. , Bjork R. A., Finley, J. R. & Linn, M. C. (2005) .Linking cognitive science to education: generation and interleaving effects.
http://www.psych.unito.it/csc/cogsci05/frame/talk/f909-richland.pdf [17.8.2005]

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. Applied Cognitive Psychology, 19, 361-374.

Rütter, T. (1973). Formen der Testaufgabe. Beck. München.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. Psychological Science, 3, 207-217.

Siegel, M. A. & Misselt, A. L. (1984). Adaptive Feedback and Review Paradigm for Computer-Based Drills. Journal of Educational Psychology, 76 (2), 310-317.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard s (1913) and Bartlett s (1932) results. Psychological Science, 3, 240-245.

Wheeler, M. A.; Ewers, M. & Buonanno, J. F. (2003) Different rates of forgetting following study versus test trials. Memory, Vol. 11 Issue 6, 571-580.

---

created: Februar 2006; Bernhard Jacobs, b.jacobs@mx.uni-saarland.de