Autor: Bernhard Jacobs, Medienzentrum der Philosophischen Fakultäten der Universität Saarbrücken
created: 1.10.2007
URL des Originals: http://www.phil.uni-sb.de/~jakobs/wwwartikel/teststudy/teststudy2.html

Die Behaltenswirksamkeit wiederholten Einprägens im Vergleich zu Computer- und selbst gesteuertem Testen mit Feedback.

Abstract

In einem klassischen Randomisierungsexperiment wurde die Hypothese geprüft, ob Testen mit Feedback langfristiges Behalten besser fördert als das erneute Einprägen der Information. Als Lehrziel diente die Zuordnung der Namen der US-Bundesstaates zu ihren Territorien auf der Landkarte. Hierbei wurden 3 verschiedene Test- bzw. Feedbackversionen mit dem Einprägen anhand einer konventionellen Landkarte verglichen. Als beste Übungsmethode erwies sich eine spezielle Testvariante, die dem Lerner eine selbst gesteuerte Testung mit Rückmeldung via clickable map erlaubte. Diese Selbsttestmethode erzielte signifikant bessere Behaltenswerte als das Einprägen auf der Basis einer konventionellen Landkarte, war den übrigen Testmethoden aber nicht signifikant überlegen. Die Computer gesteuerten Testvarianten erzielten trotz deutlich unterschiedlich aufwändigem Feedback vergleichbare Behaltenswerte und zeigten erwartungswidrig keine Behaltensvorteile gegenüber erneutem Einprägen.

Schlagworte. Aufgabenformen, test format, Übung, Practice, Feedback, Feedbackarten, Drill, clickable map, multiple choice, short answer, flashcard

Einleitung und Ausgangslage

Der "Testeffekt" (Glover 1989, Roediger & Karpicke 2006b) besagt, das Testen allein bewirke ein besseres Behalten. Wird im Anschluss an eine Lernaneignungsphase ein Test oder kein Test bearbeitet, so erzielen die Getesteten später höhere Behaltenswerte als die nicht Getesteten (Hamaker 1986). Etliche Studien belegen sogar langfristige Behaltensvorteile einer Testung, wenn die Information getestet statt erneut eingeprägt wird. (z.B.: Nungester & Duchastel 1982, Duchastel & Nungester 1984, Roediger & Karpicke 2006a, Butler & Roediger (2007)). Noch deutlichere Behaltensvorteile des Testens sind in der Regel allerdings zu erwarten, wenn sich der Testung ein Feedback anschließt (z.B. Cull 2000). Von besonderer pädagogischer Relevanz erscheinen Studien, welche den Nachweis erbringen konnten, Testen mit Feedback sei dem gezielten Studieren hinsichtlich des längerfristigen Behaltens überlegen (z.B.: LaPorte & Voss 1975, Cull 2000, Clifton 2005, Kang, McDermott & Roediger 2007, McDaniel, Anderson, Derbish, Morrisette 2007). Jacobs (2006) fand empirische Evidenz für das Lehrziel "Erlernen der Bundesstaaten der USA". In einem Wiederholungsdesign erzielten Studenten unter 3 verschiedenen Test- und Feedbackbedingungen stets numerisch höhere, in zwei Fällen signifikant bessere Behaltenswerte als unter einer Methode, welche lediglich erneutes Einprägen vorsah. Zwischen den einzelnen Testübungsmethoden (Multiple Choice, Short Answer, Covert Short Answer) konnten hingegen keine klaren Unterschiede gefunden werden.

Das hoch kontrollierte experimentelle Vorgehen von Jacobs (2006) unterscheidet sich deutlich von der pädagogischen Praxis. Wenige Studenten werden beim praktischen Lernen ihre Übungsmethoden wie in einem Wiederholungsdesign variieren. Bei allen Übungsmethoden bestimmte der Computer das Lerngeschehen. Insbesondere legte er fest, in welcher Anordnung die Items zum Einprägen oder Testen zu bearbeiten waren, wenngleich er immerhin die Bearbeitungszeit für jedes Item dem Übenden überließ. In der Praxis bestimmt in der Regel der Lerner, wie er sein Lernen organisiert. Eine wesentliche Forschungsfrage dieser Untersuchung bezog sich auf die Vermutung, der Lerner könne möglicherweise besser wissen, welche Items er wie intensiv einprägen oder testen sollte. Das Erlernen bzw. Einprägen von Staaten wird normalerweise mit Hilfe einer Landkarte im Atlas vorgenommen, bei welcher der Lerner die Namen in den

Staatsgebieten entsprechend seinen Lernbedürfnissen durchgeht. Deshalb wurde hier als Kontrollbedingung "erneutes gezieltes Studieren" das selbständige Einprägen der Bundesstaaten anhand einer Landkarte gewählt.

Besondere Testarrangements, etwa die Flashcard-Methode, können ihre theoretisch angenommenen Stärken nur unter bestimmten Bedingungen ausspielen. In dem Wiederholungsdesign von Jacobs (2006) bearbeitete der Student unter jeder Bedingung ca. 12 Items. Die Flashcard- Methode mit Response Contingent Feedback (RCF) erbrachte aber keine Vorteile gegenüber mehrfachen Wiederholungen aller Items mit einfachem KCR-Feedback. Flashcard erscheint vielleicht nur dann den Wiederholungen aller Items überlegen, wenn die fehlerhaften Items dadurch deutlich intensiver bearbeitet werden, was unter der Voraussetzung konstanter Lernzeiten bei längeren Itemlisten eher zu erwarten ist als bei kürzeren. Im Wiederholungsdesign von Jacobs (2006) beschränkte sich das Antwort abhängige Feedback auf die jeweils eingeübte Staatengruppe. Falsch angeklickte Staaten außerhalb der eingeübten Staatsgruppe konnten aus experimentellen Gründen nicht als Verwechslung rückgemeldet werden, weil dadurch Lernen für die weiteren experimentellen Bedingungen gefördert worden wäre. Beim Einüben aller USA-Bundesstaaten kann hingegen Feedback zu jedem verwechselten Staat gewährt werden. Diese und weitere Überlegungen führten letztlich zu der zwingenden Schlussfolgerung, das Experiment mit unabhängigen Gruppen durchzuführen und mit jeder Übungsmethode jeweils alle 50 Bundessstaaten der USA zu prüfen.

Probleme der Forschung, der pädagogischen Praxis, der Versuchsplanung und der Statistik

Zur strengen Überprüfung theoretischer Fragestellungen zum Effekt des Testens und des Feedbacks eignen sich bestimmte Wortlisten (Wheeler, Ewers & Buonanno, 2003), locker zusammen hängende Wortpaare (Carpenter, Pashler & Vul. (2006), Eskimo-Vokabeln (Carrier & Pashler 1992) oder obscure objects bzw. obscure facts (Pashler, Rohrer, Cepeda & Carpenter 2007) sicherlich besser als die Bundesstaaten der USA. Wenn man dann schon eine Landkarte verwendet, was den experimentellen Aufbau allein schon verkompliziert, dann hätte man aus der Sicht strenger Laborforschung wenigstens die Landkarte frei erfinden müssen (Carpenter & Pashler (2007)), um sicher zu stellen, dass jeder Lerner mit "Nullwissen" zum Experiment antritt. Das rigorose Laborexperiment selbst mit praxisfernen, künstlich ausgewählten Reizen hat sicher seine Berechtigung, weil Störfaktoren so besser zu isolieren sind, die Prüfung stringenter auf eine Hypothese ausgerichtet und der statistische Nachweis leichter erbracht werden können. Hierbei wurde der Lernvorteil des Testens mit Feedback gegenüber dem gezielten Einprägen aber bereits häufig nachgewiesen.

Es ist darüber hinaus aber wichtig, die Tragweite der dort gefundenen Effekte in realitätsnäheren Schulumwelten zu überprüfen, um die Anwendbarkeit und den pädagogischen Nutzen besser abschätzen zu können (=Problem der externen Validität). So wertvoll theoretisches Grundlagenwissen auch sein mag, es bezieht sich häufig auf ideale Situationen, die in schulischen Umwelten in dieser Form gar nicht bzw. mehr oder weniger stark vorkommen. Die erfolgreiche Anwendbarkeit theoretischen Wissens ist somit stets an bestimmte Zusatzbedingungen gebunden. Mittlerweile konnte die Überlegenheit des Testens mit Feedback gegenüber gezieltem Studieren aber auch in realistischen Schulumwelten belegt werden (z.B. McDaniel et al. 2007).

Die meisten Studenten kennen zumindest einige Bundesstaaten der USA, (etwa Florida, California, Texas) und bereits vor jedem Lernen sind schon Vorwissensunterschiede zu vermuten. Das ist bei vielen praktischen Lerngebieten so und erweist sich lediglich bei einigen Grundlagenexperimenten als unerwünscht. Bisherigen Erfahrungen zufolge ergaben Testungen nach der Lernaneignungsphase, d.h. vor der experimentellen Intervention, mittlere Erfolgsquoten, aber eine ungewöhnlich hohe Personenstreuung. Unter diesen Bedingungen benötigt man sehr viele Proban-

den für ein Treatment, um experimentelle Effekte durch ein klassisches Randomisierungsexperiment ohne Vortest nachweisen zu können. Auf ein sensitiveres Design mit einem generellen Vortest, welches den Effekt der experimentellen Variation als Posttest-Vortest-Differenz hätte ermöglichen können, musste jedoch verzichtet werden, da der Vortest ja einen Testeffekt bewirkt und dann die Auswirkungen reinen ausschließlichen Einprägens nicht mehr erfassbar sind. Alle mir bekannten Studien, welche einen Vorteil des Testens mit Feedback gegenüber gezieltem Einprägen nachweisen konnten, basieren auf einem Wiederholungsdesign. Die statistisch gesicherten Befunde liegen im niedrigen bis höchstens mittleren Effektstärkebereich.

Bei vermuteten Unterschieden von höchstens ca. 0.5 Effektstärke sollten mindestens 50 Probanden pro Treatment verfügbar sein, um eine akzeptable Chance für einen statistischen Nachweis mit unabhängigen Gruppen erbringen zu können. Die Teststärke betrüge dann für die einseitige Testung ca. .80 (siehe z.B.: Lenth, 2006 Java applets for power and sample size : two sample t-test.). Da mir keine Forschungsgelder zur Verfügung stehen und ein entsprechender Antrag für Vpn-Gelder abgelehnt wurde, habe ich mich zunächst entschlossen, das Experiment mit freiwilligen Probanden durchzuführen. Nachdem ich nun aber schon über 1,5 Jahre die Sache verfolge, bei ca. knapp der Hälfte der erforderlichen Probandenanzahl angekommen bin, selbst damit rechnen muss, dass durch die Einführung neuer Browserversionen die zur Zeit noch funktionsfähigen programmierten Experimentalbedingungen in Zukunft nicht mehr lauffähig sind, bin ich zum Entschluss gelangt, die Untersuchung mit allen damit in Kauf zunehmenden Konsequenzen abzuschließen. **Die zu geringe Probandenanzahl wirkt sich deshalb besonders negativ aus, weil die statistische Prüfung nun die Form eines Glücksspiels annimmt.** Aber manche Teile des Experimentes sind recht gut gelungen und einige Ergebnisse erscheinen teilweise interessant, so dass es mir nicht ganz nutzlos vorkommt, das Forschungsvorgehen und die Ergebnisse in diesem Bericht zu dokumentieren. Um den Betafehler wenigstens teilweise einzuschränken, habe ich das Signifikanzniveau auf 10% festgesetzt. Damit dürfte die Testpower wenigstens etwas höher als 50% ausfallen. Darüber hinaus will ich notgedrungen eine pädagogisch pragmatische Sicht verfolgen: "Wenn sich bei Gruppenstärken, die in etwa einer Klassengröße entsprechen, keine deutlichen Unterschiede zeigen, dann sind die erhofften Effekte, auch wenn sie tatsächlich existieren sollten, zu dünn, um sich für das hier verfolgte Lehrziel als ernsthafte pädagogische Alternative in der Praxis aufzudrängen." Das ist freilich eine sehr strenge Forderung, die man selten in Erwägung zöge, wenn es beispielsweise um die diagnostische Effizienz einer medizinischen Vorsorge, den Nachweis der Wirksamkeit einer neuen Operationsmethode, oder den statistischen Beleg für den Vorteil eines neuen Medikamentes gehen würde.

Zielsetzung der Untersuchung

Die Studie dient zum einen Replikationszwecken, indem wiederum überprüft werden soll, ob Übungsmethoden, die ein Testen mit Feedback beinhalten, Behaltensvorteile gegenüber einer Übungsmethode aufweisen, die den Schwerpunkt auf die Darbietung der Information zum erneuten Einprägen legt. Da hierbei einige Bedingungen in modifizierter Weise zum Einsatz kommen, beschränkt sich die Replikation nicht auf eine triviale Wiederholung. Unter Lernerfolg wird eher ein relativ langfristiges Behalten von ca. einer Woche nach der Übung verstanden. Denn bisherige Studien belegen vornehmlich Testvorteile im Hinblick auf längerfristiges Behalten (z.B. Roediger Karpicke 2006a), was sich in der Studie von Jacobs (2006) insofern bestätigte, als im unmittelbaren Nachtest keinerlei Treatmentunterschiede zu finden waren, die Testmethoden aber nach 3 bis 7 Tagen höhere Behaltenswerte erzielten als erneutes Studieren.

Zum andern interessiert, welche Auswirkungen die Übungssteuerung auf das Behalten ausübt. Hierbei macht zum einen der Computer Vorgaben, in dem er die zu bearbeitenden Itemreihenfolge festlegt, zum anderen muss der Lerner das Übungsgeschehen selbst bestimmen. Selbstbestimmte Lernmethoden entsprechen eher der üblichen pädagogischen Praxis und allein schon deshalb ist ihr Übungseffekt wichtig zu erfahren. Zudem verband sich mit der Variante Selbsttesten die Hoffnung auf eine verbesserte Lernmethode, wenngleich dieser Vorteil nur bei rationaler Anwendung zu erwarten ist. Im Idealfall testet der Lerner relativ gut beherrschte Staaten sehr schnell ab, nimmt sich mehr Zeit für die Enkodierung problematischer Items und überprüft deren Kenntnis durch häufige Testung. Schließlich sollte überprüft werden, ob eine aufwändigere Feedbackprozedur, die Antwort abhängiges Feedback umfasst sowie die Flashcardmethode mit einer stärkeren Vorgabe fehlerhafter Items unter günstigeren Anwendungsbedingungen als bei

Jacobs (2006) mehr Lernerfolg erzielt als sehr sparsames Covert Short Answer mit der schlichten Rückmeldung der korrekten Antwort.

Die empirische Untersuchung

Versuchspersonen

An der Untersuchung nahmen insgesamt 99 Studierende der Universität des Saarlandes teil. Das Durchschnittsalter betrug 26 Jahre. Lehramtsstudierende und Studierende der Erziehungswissenschaften, die an Seminaren des Verfassers oder von Herrn Paulus teilnahmen, sowie Studierende der Informationswissenschaften 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 Dr. Christoph Paulus sowie den Tutoren Armella-Lucia Vella und Nina Michaltzik für die Organisation von Probanden, sowie allen beteiligten Probanden, die ca. eine Stunde ihrer Zeit unentgeltlich der Wissenschaft zur Verfügung stellten.

Versuchsablauf

Abbildung 1 : Versuchsablauf

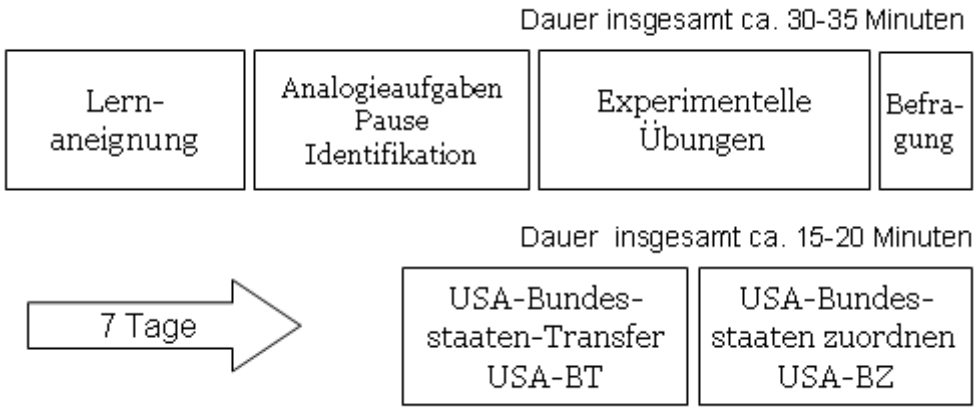


Abbildung 1 gliedert den Versuchsablauf. in zwei 2 Sitzungstermine. Der erste Termin diente der Lernaneignung und den experimentellen Übungsmethoden, der zweite Termin den Nachmessungen zum Zwecke der Überprüfung des Behaltens. In der Lernaneignungsphase wurden die Bundesstaaten der USA zweimal zum Einprägen präsentiert. Die Analogieaufgaben hatten lediglich die Funktion, zwischen Lernaneignung und experimenteller Übung einen Zeitpuffer zu schaffen, um das Übungsgeschehen wenigstens etwas zu verteilen. Die Lernaneignung entspricht exakt dem Vorgehen bei Jacobs (2006). Um die ökologische Validität der Übung zu erhöhen, stand in der Anweisung unmittelbar vor der experimentellen Übung abweichend zu Jacobs (2006) der Hinweis, der Lerner könne sich auch Notizen machen. Der Untersuchungsleiter ermutigte die Studenten bei Unklarheiten, so zu lernen, wie sie es für sinnvoll erachteten. Die Studierenden wurden nach Zufall auf 4 verschiedene Übungsbedingungen aufgeteilt. Die experimentelle Phase dauerte für alle Treatmentvarianten exakt 12 Minuten. Sie wurde gegenüber Jacobs 2006 insgesamt um 4 Minuten gekürzt, da es einfacher erschien, nach einer Lernmethode zu üben als bei Jacobs 2006. Dort bearbeitete jeder Student in einem Wiederholungsdesign 4 verschiedene Lernmethoden. Nach der experimentellen Übung folgte eine kurze Befragung zur Einschätzung der Gesamtübung. Am Ende wurde den Studenten mitgeteilt, in einer Woche fände ein weiterer

Untersuchungstermin statt, aber keinerlei Information gegeben, was dort erfasst werden sollte. Die Nachtests fanden exakt eine Woche später zur gleichen Zeit statt.

Die experimentellen Übungsmethoden

Tabelle 1 beschreibt die experimentellen Übungsmethoden und verweist auf die entsprechenden Programme, welche einen genauen Einblick in den Ablauf der jeweiligen Bedingung erlauben. Die verwendete Grafik stammt aus: http://www.jayzeebear.com/map/usa.html [18.5.2005]. Ich danke dem Konstrukteur der Grafik für die Erlaubnis, diese für wissenschaftliche Zwecke verwenden zu dürfen. Bildschirmkopien der Programme befinden sich im Anhang.

Tabelle 1: Beschreibung der Übungsmethoden

Übungsmethode: **Userkennung: Alabama; Passwort: Alabama**	
Erneutes Studieren mit der Landkarte: Study only (**SO**)**:** Auf dem Bildschirm erscheint die Karte der USA-Bundesstaaten einschließlich der Namen der Bundesstaaten in den entsprechenden Territorien. Der Lerner hat die Aufgabe, mit einer ihm sinnvoll erscheinenden Strategie die Bundesstaaten einzuprägen.	**Studieren via konven- ioneller Land- karte**
Selbsttesten mit CSA (Selbsttesten) Auf dem Bildschirm ist eine Karte mit allen Bundesstaaten der USA ohne Namen zu sehen. Der Lerner soll ein beliebiges Staatsgebiet auswählen und versuchen, sich an den Namen zu erinnern. Durch Anklicken des entsprechenden Gebietes kann er den Namen des Staates einsehen. Zusätzlich hat der Lerner die Option jederzeit in einen SO-Modus zu wechseln und auf einen Blick alle Staatsnamen in den entsprechenden Gebieten zu sehen.	**Selbsttesten**
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 45 Sekunden alle Staatsnamen im entsprechenden Staatsgebiet.	**Covert Short An- swer mit KCR- Feedback:**
Multiple Choice Test mit KOR+KCR+ 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 30 Sekunden alle zuvor falsch beantworteten und verwechselten Staatsnamen im entsprechenden Staatsgebiet. (Genaueres siehe Anhang)	**Multiple Choice Test mit KOR+KCR+ RCF Feedback+ Flashcard**

*****Der echte Versuch lief im 15 bzw. 17 Zoll-Vollbildmodus ohne Menu-, Symbol-, oder Statuszeilen.
Die Programme dienen nur als Demonstration. Das JavaScript-Programm ist unverständlich und beinhaltet etliche Eigentümlichkeiten. Neuere Browserversionen oder Bildschirme können die korrekte Darstellung beeinträchtigen.

Durch die Vorgabe der Landkarte unter Study Only werden vom Lernmedium lediglich die relevanten Informationen zum Einprägen zur Verfügung gestellt. Weil es dem Studenten aber überlassen bleibt, diese nach seinem Belieben für Lernzwecke zu nutzen, kann nicht völlig ausgeschlossen werden, dass er sich teilweise selbst testet, wenngleich dies am Computer schwerer

sein dürfte als z.B. bei einer Landkarte im Atlas. Einige Probanden nutzten Papier für Notizen und malten teilweise einzelne Staaten auf, was auch als Grundlage einer Selbsttestung dienen kann. Alle Testvarianten enthalten neben klassischem Feedback zeitweise auch eine spezielle, ausschließliche Informationsvariante, welche meist alle Staaten in den Gebieten gleichzeitig darstellt. Nur bei der MC-Übungsvariante beschränkte sich diese Informationsvariante auf die fehlerhaften Staatsnamen, ein Verfahren, was in ähnlicher Weise bereits bei Thomson, Wenger & Bartling (1978) zur Anwendung kam. Eine realistische Schätzung der tatsächlichen Nutzung aller Testvarianten lässt vermuten, dass die echten Testungen mit normalem Feedback mindestens 90% der Gesamtzeit dieser Übungsvarianten ausmachten. Unter Study only und Selbsttesten bestimmt der Lerner das Übungsvorgehen, bei CSA und MC organisiert der Computer die Itemvorgaben.

Abhängige Variablen

A) **USA-Bundesstaaten-zuordnen-können (**USA-**BZ)**
Im USA-BZ wurde der Name eines Bundesstaates vorgegeben und der Proband hatte die Aufgabe, das entsprechende Territorium auf der Landkarte mit der Maus anzuklicken. (analog der Übungsmethode MC, jedoch ohne Rückmeldungen). Dieser auch unter dem Namen clickable map bekannte Aufgabentyp entspricht nach Rütter (1993) einer MC-Aufgabenform, weil die korrekte Antwort in der Aufgabenstellung enthalten ist. Alle 50 Bundesstaaten der USA wurden für jeden Probanden stets in zufälliger Reihenfolge dargeboten. In einer früheren Untersuchung erzielte der USA-BZ, dort MC-Test genannt, eine interne Konsistenz von α=.91, eine Retestreliabilität nach 3 bis 7 Tagen von ca. .75, sowie eine Korrelation von r=.88 mit einem Short Answer Test, der bei Vorgabe eines Territoriums die schriftliche Eingabe des Staatsnamens verlangte.

b) **USA-Bundesstaaten Transfer Test (**USA-**BT)**
Im Transfertest wurden Aspekte geprüft, die in den experimentellen Übungen niemals explizit eingeübt oder bei den Testbedingungen gezielt getestet wurden, sich beim Einüben aber mehr oder weniger implizit anbieten könnten oder sich aus dem Wissen der Zuordnung ableiten lassen. Je nach Memorierungstechnik oder Verarbeitungstiefe des Einzelnen dienen sie als potentielle Orientierungspunkte. Häufig erfordern die Fragen ein Operieren mit den Bundesstaaten. Mit der Konstruktion des Transfer-Tests war auch die Hoffnung verbunden, vielleicht einige Unterschiede zwischen den experimentellen Übungen zu Gunsten der selbst bestimmten Übungsmethoden zu finden, die weniger stures Einpauken nahe legen und mehr Gewicht auf die Beziehungen zwischen den Ländern bzw. der Einordnung eines Staates in das Gesamtgebilde der USA fördern könnten.

Zum BT gehören etwa Fragen nach Besonderheiten bestimmter Staaten wie das Erinnern an die Flächenform [z.B. stark rechteckig oder stark von der Rechtecksform abweichend], das Erkennen bestimmter markanter Staatsumrisse, [etwa Louisiana, Idaho], das Wissen von Relationen der Staaten untereinander [z.B. flächenmäßig eher größer oder kleiner, Position des Staates innerhalb der USA eher zentral oder peripher], das Wissen über die relative geographische Lage eines Staates [z.B. im Südosten, Nordwesten] oder mehrerer Staaten untereinander [etwa Staat zwischen Ohio und Illinois?] oder die Positionierung von Staatsanhäufungen [Im Nordosten mehr Staaten als im Südosten ?]. In vielen Fällen ist ein solides Wissen im BZ Voraussetzung für gute Leistungen im BT. Der BT besteht letztlich auch 9 kleineren, auf verschiedenen Aufgabentypen [Multiple-Response-MC, True/false, Zuordnung, Short Answer] basierenden Aufgabenkomplexen bzw. kleinen Untertests, die stets mehrere Fragen oder Antworten zu einem Themenkomplex beinhalten und deren Erfolgsquoten aufaddiert den entsprechenden Testwert im BT ergeben.

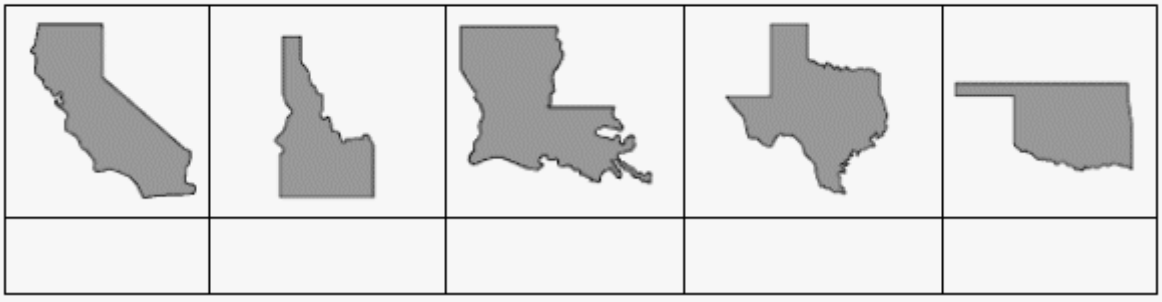
Aufgabenbeispiele aus dem USA-BT:

Welche **3** der nachfolgenden 7 Bundesstaaten liegen ziemlich zentral in der Mitte der USA

1) Colorado
2) Nebraska
3) Alabama
4) Kansas
5) Idaho
6) Tennessee
7) Indiana

[3 Antworten sind richtig]

Sie sehen unten 5 Bundesstaaten, die nicht Ihrer tatsächlichen Größe nach vergleichbar sind. Entscheidend ist vielmehr nur die Form.
Wie heißen diese Staaten ?



Alle Aufgaben des USA-BT waren auf einer Seite für alle Probanden in der gleichen Reihenfolge angeordnet. Alle Probanden beantworteten zunächst den BT und anschließend den BZ.

Reliabilität der abhängigen Variablen

5% der Probanden erzielten im Test "Bundesstaaten der USA zuordnen können" Erfolgsquoten von kleiner 15% bzw. größer 90 Prozent korrekter Lösungen. Um die Reliabilität der abhängigen Variablen nicht künstlich zu erhöhen, wurden diese Probanden aus der Reliabilitätsanalyse ausgeschlossen.

Tabelle 1: Cronbachs α und Interkorrelation der Tests (N jeweils 94):

Bundesstaaten zuordnen können	(BZ)	α = .93
Bundesstaaten Transfer Test	(BT)	α = .75
Korrelation zwischen BZ und BT		r = .82

Die Reliabilitäten fallen insgesamt zufrieden stellend aus. Die Korrelation zwischen BZ und BT unterstützt die These, die Kenntnis der Bundesstaatenzuordnung sei eine wesentliche Stütze für den Transfer. Sie fällt aber etwas zu hoch aus, um darauf hoffen zu können, man habe mit dem Transfertest auch Faktoren erfasst, die weniger von der reinen Kenntnis der Bundesstaaten ab-

hängen, sondern auch durch die verschiedenen Übungsmethoden unterschiedlich beeinflusst werden könnten.

Ergebnisse

Überprüfung der Randomisierung

Die Effizienz der Randomisierung wurde durch einen Vergleich der Gruppen hinsichtlich des erfragten Abiturnotendurchschnitts und des Alters überprüft. Hierbei ergaben sich signifikant schlechtere Abiturdurchschnittsnoten der Bedingung MC (2.6) gegenüber allen anderen Bedingungen (alle ca. 2.3), sowie ein signifikant höheres Alter der Gruppe Selbsttesten gegenüber den Gruppen CSA und MC. Die gefundenen Unterschiede belasten allerdings kaum die interne Validität, da der Abiturnotendurchschnitt (im Gegensatz zu der Studie von Jacobs 2006) sowie das Alter insignifikant mit der abhängigen Variablen USA-BZ korrelieren (BZ, Abi: r = -.17; BZ, Alter r= .08).

Einige Lernprozessdaten

Eine zeitlich beschränkte Übung muss bei allen Übungsvarianten faire Lernchancen gewähren, die zumindest hinreichende Lernmöglichkeiten bieten, alle Items gründlich zu bearbeiten. Während diese Forderung für die Bedingung Study only (Landkarte) auf Grund von Plausibilitätsüberlegungen lediglich theoretisch angenommen werden kann, da deren Bearbeitung keinerlei Computeraktionen erforderte und somit keine Messoperationen zuließ, wurden bei den übrigen Übungsmethoden während der Übung einige Daten erhoben, die gewisse Rückschlüsse auf die Aufgabenbearbeitung zulassen. Bei der Methode CSA gab der Computer alle Items stets in zufälliger Reihenfolge vor und speicherte nur die Anzahl der insgesamt bearbeiteten Aufgaben. Praktisch alle Probanden haben unter Übungsmethode CSA wenigstens einmal alle Items zum Testen angefordert. Der Durchschnitt liegt etwas über zwei vollständigen Übungsdurchgängen.

Beim Selbsttesten konnte der Lerner im Testmodus frei bestimmen, welche Staaten (Items) er in welcher Reihenfolge und Häufigkeit auch immer via Covert Short Answer mit nachfolgendem KCR-Feedback anfordert. Die Erhöhung der Itemhäufigkeit um ein weiteres Item setzte allerdings voraus, dass der Lerner zunächst mindestens einen anderen Staat dazwischen anwählen musste. D.h. mehrfaches Anklicken desselben Staates unmittelbar hintereinander wurde als eine Itembearbeitung gewertet. Im Durchschnitt bearbeiteten die Probanden ca. 200 Items. Es wäre möglich, den Lernweg der Studenten nachzuvollziehen, um die speziellen Strategien und Vorgehensweisen der einzelnen Probanden in Erfahrung zu bringen, worauf hier aus Zeitgründen jedoch verzichtet wird. Beim Selbsttesten konnte der Lerner weiterhin jederzeit von einem Testmodus in einen Studiermodus (=Study only (Landkarte)) sowie vom Studiermodus in den Testmodus wechseln. Im Durchschnitt wechselten die Probanden ca. zweimal vom Test- in den Studiermodus. Im arithmetischen Durchschnitt verweilten die Lerner im Studiermodus etwas mehr als eine Minute, im Median allerdings nur 17 Sekunden. Ca 40% aller Studenten wählten die Studiervariante praktisch gar nicht aus. Nur ca. ein Drittel der Studenten nutzte die Studieroption überhaupt in einem nennenswerten Zeitausmaß von mindestens einer Minute. Betrachtet man die realisierte Nutzungszeit der beiden Übungsoptionen innerhalb des Selbsttestens als objektiven Verhaltenstest, so haben die Studenten freiwillig mit ca. 90% der verfügbaren Übungszeit das Testen mit Feedback ganz eindeutig gegenüber dem Studieren mittels konventioneller Landkarte vorgezogen.

Ein Flashcarddurchgang unter der MC-Bedingung ist erst dann beendet, wenn der Proband jedes Items einmal richtig beantwortet hat. Danach sah er für eine halbe Minute alle falschen und ver-

wechselten Staatsnamen in den entsprechenden Staatsterritorien. Eine derartige Übungskonzentration auf die kritischen Items war nur bei dieser Variante möglich, weil hier in der Übung objektive Testdaten verfügbar waren. Anschließend begann der nächste Flashcarddurchgang. Dieser Vorgang wiederholte sich bis zum Übungsabbruch. Beim Überschreiten der zulässigen Bearbeitungszeit von 12 Minuten wurde ein laufender Flashcarddurchgang somit abgebrochen. Bei konstanter Arbeitsgeschwindigkeit dauert ein Flashcarddurchgang umso länger, je mehr Fehler ein Proband macht. Die individuelle Bearbeitungsgeschwindigkeit war aber dem einzelnen überlassen. Von 29 Probanden der MC-Bedingung absolvierten

20 mindestens einen vollständigen Flashcard-durchgang,
7 mindestens 2 Flashcard-durchgänge,
2 mindestens 3 Flashcard-durchgänge,

Die Studenten bearbeiteten im Durchschnitt 114 Items. Basierte ein Fehler auf dem Anklicken eines falschen Staates, so liegt eine Verwechslung vor. Die entsprechenden Daten für die Verwechslungen lauten M=42, s=29. Die MC-Übung muss trotz der Zeitbeschränkung insgesamt als faires Übungsangebot gewertet werden, da jeder Proband mindestens einmal alle Items richtig beantwortete. Die verwendete Flashcard-Methode erscheint zudem theoretisch sinnvoll, da zu Beginn ca. 50% aller Items falsch gelöst wurden, im Fehlerfalle KCR-Feedback angeboten sowie im weiteren Verlauf die falschen Items verstärkt zur Testbearbeitung und Feedbacknutzung vorgelegt wurden. Die relativ hohe Anzahl von 42 Verwechslungen belegt rein theoretisch die potentielle Effektivität des response contingent feedback. Denn bei 37% aller Aufgabenbearbeitungen wurde die Chance angeboten, sich neben dem geforderten Staatsnamen auch noch den zutreffenden Namen des verwechselten Staates einzuprägen. Die hohe Zahl der Verwechslungen lässt vermuten, dass die Probanden häufig geraten haben und im Falle hoher Unsicherheit seltener die Option "weiß nicht" gewählt haben.

Vergleich der Itemhäufigkeiten

Bei konstanter Übungszeit dürften die einzelnen Übungsmethoden unterschiedlich viele Aufgabenbearbeitungen begünstigen bzw. nahe legen. Nur die Testübungsmethoden ermöglichten messbare Aktionen zur Anwahl eines bestimmten Bundesstaates. Tabelle 2 stellt die Anzahl der bearbeiteten Items für alle 3 Testübungsmethoden dar.

Tabelle 2: Anzahl der bearbeiteten Items (N ca. 26 pro Gruppe)

	M	s
Selbsttesten	198	66
CSA	122	49
MC	114	35

Wie die Mittelwerte in Tabelle 2 anschaulich aufzeigen, haben die Probanden unter der Übungsmethode Selbsttesten deutlich mehr Items bearbeitet als die Probanden unter den restlichen Testübungsgruppen. Die praktisch sehr bedeutsamen Unterschiede zwischen Selbsttesten mit CSA sowie mit MC sind nach t-Test hochsignifikant, während sich die Bearbeitungshäufigkeiten von CSA und MC nicht voneinander unterscheiden.

Unmittelbarer Lernerfolg nach der Lernaneignungsphase

Es war beabsichtigt, durch die Lernaneignungsphase ca. 50% Lernerfolg zu erzielen, um zum einen ein gewisses Lernfundament zu legen, zum andern aber noch Spielraum für die experimentellen Übungen zu ermöglichen. Die Übungsmethode MC beinhaltet selbst eine MC-Testung. Da die Probanden nach Zufall den Treatmentbedingungen zugeordnet wurden, erlaubt die erste Testung der Items unter der MC-Bedingung für alle beteiligten Probanden dieser Untersuchung eine

grobe Schätzung des Lernerfolgs ca. 15 Minuten nach der Lernaneignungsphase. Diese Schätzung dürfte etwas erhöht ausfallen, da sich der Testung jedes Items ein KCR- und RCF-Feedback anschloss und somit während des Testens teilweise auch hinzugelernt wurde bzw. auch werden sollte. Die Probanden ordneten im ersten Flashcarddurchgang der MC-Übung ca. 50% aller Bundesstaaten korrekt zu. (M= 49.7%, s= 19,7%; N=29). Dieses Ergebnis entspricht fast exakt den früheren Erfahrungen (Jacobs 2006). Die ursprüngliche Hoffnung, die experimentellen Übungen mit einer durchschnittlichen Lösungswahrscheinlichkeit von .5 zu beginnen, hat sich demnach in idealer Weise erfüllt.

Auf eine Erhebung des Lernerfolgs unmittelbar nach den experimentellen Übungen wurde verzichtet, da der langfristiger Lernerfolg von Interesse war und die unmittelbare Testung im Anschluss an die Übung einen deutlichen Einfluss auf den langfristigen Lernerfolg gehabt hätte. (siehe: Führt das Testen nach einer Übung zu einem verbesserten langfristigen Behalten ?) Als realistische Schätzung könnte man den Ergebnissen von Jacobs (2006) zufolge ca.70% korrekter Lösungen annehmen. Wegen dem dann einsetzenden Vergessen sind nach einer Woche natürlich deutlich weniger korrekte Lösungen zu erwarten und hier stellt sich ja die Frage, mit welcher Übungsmethode man am meisten behält.

Behaltensdaten, eine Woche nach der Übung

In die Berechnungen der Behaltenswerte gingen nur solche Personen ein, die eindeutig einer bestimmten Bedingung zugeordnet werden konnten. Ein Proband aus Bedingung MC wurde wegen extrem schwacher Leistung (12%), ein anderer Proband aus Bedingung Selbsttesten wegen extrem guter Leistung (100%) aus der Analyse ausgeschlossen. Ungleiche Häufigkeiten für die einzelnen Gruppen basieren hauptsächlich auf Personenschwund, da etliche Probanden nicht mehr zum Nachtest angetreten waren und einige Probanden falsche Angaben zur Identifikation machten.

Es waren eindeutig höhere Behaltensleistungen aller Testübungsmethoden gegenüber Study only erwartet worden. Einfaktorielle VA-analysen mit allen Übungsvarianten als Faktorstufen erbrachten für den BZ mit F (3,88) = 1.07; p=0.37 und für den BT mit F(3,88) = 0,9; p=0.44 jedoch eindeutig insignifikante Ergebnisse. Tabelle 3 zeigt die relevanten Daten für alle experimentellen Gruppen

Tabelle 3: Prozentsatz korrekter Lösungen im Posttest (eine Woche nach der Übung)

		Bundesstaaten zuordnen USA-BZ		**Bundesstaaten Tranfer USA-BT**	
	N	M	s	M	s
Study only (Landkarte)	23	40,3	19,1	51,8	11,6
Selbsttesten (CSA)	18	52,2	23,6	56,2	16,0
Covert Short Answer	28	43,6	21,7	49,2	16,1
Multiple Choice	23	44,5	22,7	50,0	15,4

Für den BZ deuten sich numerisch einige schwache Vorteile für die Testversionen an. Nachfolgende Tests ergaben aber lediglich für den BZ einen signifikanten Vorteil des Selbsttestens gegenüber Study-only. Der t-Tests zur Überprüfung der Mittelwertsunterschiede zwischen Selbsttesten und Study-only unterbot für den BZ mit t(39) = 1.8, p=0.04; einseitig) das geforderte Signifikanzniveau von 10%. Dieser Unterschied entspricht einer Effektstärke von d =.55 zugunsten des Selbsttestens. Zumindest bei einer Testvariante hat sich der erwartete direkte Behaltensvor-

teil gegenüber ausschließlichem Einprägen somit signifikant bestätigen lassen. Der BZ-Testvorteil beim Selbsttesten war allerdings nicht massiv genug, um sich auch im Transfertest gegenüber Study only durchzusetzen. Der numerische Vorteil des Selbsttestens gegenüber den restlichen Testübungsmethoden konnte statistisch nicht gesichert werden.

Subjektive Einschätzungen

Am Ende der Übung sollten die Probanden die gesamte Übung an Hand einiger, jeweils 7 Skalenwerte umfassender, bipolar skalierter Fragen hinsichtlich Interessensanregung, Lernmotivierung, Lernunterstützung und Unterrichtsqualität einschätzen. Eine Faktorenanalyse aller Items ergab eine höchstens zweifaktorielle Lösung. Der wichtige erste Faktor kann als Wertschätzung des Übungsprogamms bezeichnet werden. Die Wertschätzung des Übungsprogramms umfasst 7 Items "z.B. Das Übungsprogramm verbessert das Einprägen deutlich" oder "Die Übung motiviert zum konzentrierten Einprägen" und erzielte ein α von .91. Der zweite Faktor "Eigene Lernsteuerung" (3 Items, α = .74) beschreibt am ehesten so etwas wie die wahrgenommene Entscheidungsfreiheit sowie die selbstbestimmbare Eigenaktivität: "z.B. Das Programm fördert den Entscheidungsspielraum für eigene Lerntätigkeit".

Es war zum einen erwartet worden, dass die Wertschätzung des Programms bei allen Testversionen höher als unter Study only ausfällt. Da Study only und Selbsttesten deutlich mehr eigenen Freiheitsspielraum gewähren, waren hier auch höhere Einschätzungen vermutet worden als unter CSA und MC.

Tabelle 4: Subjektive Einschätzungen der Übungsmethoden

		Wertschätzug des Programms		**Freie Lernsteuerung**	
	N	M	s	M	s
Study only (Atlas)	29	4,7	1,4	4,2	1,4
Selbsttesten	24	4,8	1,4	3,9	1,4
Covert Short Answer	31	4,8	1,1	4,0	1,4
Multiple Choice	28	5,1	0,8	4,4	1,3

Auf eine Signifikanztestung der in Tabelle 4 gezeigten Mittelwerte wird verzichtet, da die deskriptiven Ergebnisse die überwiegend vergleichbaren Einschätzungen für alle Bedingungen hinlänglich verdeutlichen. Die Daten können die zuvor gehegten Erwartungen insgesamt nicht bestätigen, wenngleich sich bei der Wertschätzung der erwartete Vorteil zugunsten von MC zumindest numerisch andeutet. Aber lediglich auf Itemniveau ließ sich ein signifikanter Vorteil von MC gegen study only bei dem zentralen Wertschätzungsitem " Ich bewerte die Qualität des Übungsprogramms insgesamt als hervorragend" signifikant bestätigen ($t(56)=1.87$; p einseitig =0.034; Effektstärke=0.49).

Zu bedenken bleibt, dass hier die gesamte Übung (einschließlich Lernaneignung und Analogieaufgaben) bewertet wurde und die experimentellen Bedingungen nur einen Teil, ca. ein Drittel der Zeit, ausmachten, was die Entdeckung von Treatmentunterschieden natürlich erschwert. In der früheren Studie auf der Basis eines Wiederholungsdesigns (Jacobs 2006) bewerteten die Studenten ausschließlich die experimentellen Varianten und diese in einem direkten Vergleich miteinander. Dort konnten ganz klare Treatmentunterschiede in der Bewertung der Lerneffektivität zugunsten der Testversionen gegenüber Study only gesichert werden.

Zusammenfassung und Diskussion

Die Untersuchung unterscheidet sich in einigen Punkten von den meisten mir bekannten Experimenten mit der Fragestellung, ob Testen mit Feedback einen höheren Lernerfolg bewirkt als erneutes gezielten Studieren, z.B. durch die gründliche Lernaneignung, eine ziemlich umfangreiche experimentelle Übungsphase und einen größeren Freiheitsspielraum beim Lernen. In der Lernaneignungsphase wurden alle Items zweimal zum Einprägen präsentiert. Jedes Treatment begann somit mit zwei Studierdurchgängen (S). Die Anzahl der dann folgenden Itembearbeitungen war abweichend von den typischen Experimenten dem Lerntempo der einzelnen Studenten überlassen und variierte deshalb bei der konstanter Übungszeit von 12 Minuten. Im Durchschnitt bearbeiteten die Studierenden die Items in der Experimentalphase jedoch mindestens 2 mal. Daraus ergibt sich angenähert die in Tabelle 5 angedeutete Abfolge der Studierphasen.

Tabelle 5: Deutung der Studierphasen bei den experimentellen Gruppen

Study only (Landkarte)	S	S	S	S	...
Selbsttesten	S	S	TS	TS	...
Covert Short Answer	S	S	TS	TS	...
Multiple Choice	S	S	TS	TS	...

S = Studieren, Einprägen
TS= Testen mit Feedback
... individuelle Variationen der Itemhäufigkeit bei der letzten Phase

Testen mit Feedback nicht durchgängig besser als erneutes Einprägen

Die Hypothese, alle Testmethoden bewirkten ein besseres Behalten als Study only, konnte definitiv nicht bestätigt werden. Die numerischen Werte von MC und CSA liegen zwar etwas höher als die von Study only, die Unterschiede bewegen sich jedoch im Bereich des Zufalls. Wegen relativ hohem Betafehler fällt es andererseits recht schwer, zu behaupten, nun einen fundierten Nachweis geliefert zu haben, Testen mit Feedback würde Behalten genau so gut oder schlecht fördern wie erneute Informationsaufnahme. Allerdings lassen die Daten den Schluss zu, keine großen Unterschiede zwischen den Treatments zu vermuten, was ja auch die bisherigen Befunde bestätigten. Bei Jacobs (2006) war die CSA-Methode dem ausschließlichen Studieren signifikant überlegen. Es bleibt schwer einzuschätzen, ob die veränderte Form des gezielten Studierens (hier SO als einfache Landkarte) mit dafür verantwortlich ist, dass der Testeffekt keine hinreichende Wirkung zeigte. Möglicherweise kann man besser Informationen behalten, wenn man selbst bestimmen kann, welche Items man wie oft und wie lange einprägt. Diese Vermutung scheint allerdings ziemlich unwahrscheinlich: Denn Carpenter & Pashler (2007) berichten - allerdings auf der Basis eines Wiederholungsdesigns - auch signifikante Behaltensvorteile einer CSA-Variante [dort covert retrieval genannt] gegenüber dem ausschließlichen Studieren anhand einer Landkarte. Außerdem entsprechen die Testwerte von SO und MC von den absoluten Erfolgsquoten sehr gut den früheren Befunden von Jacobs (2006), obwohl die Übung etwas kürzer und das Retentionsintervall hier etwas länger ausfallen. [SO: 40.3 vs. 40.7; MC: 44.5 vs. 46]

Neue Übungsmethoden müssen sinnvoll angewendet werden, um ihr Leistungspotential auch auszuschöpfen. Alle Testmethoden können beispielsweise sehr leicht als Study Only-Variante missbraucht werden, wenn der Lerner ohne viel Nachdenken einfach das Feedback einfordert. Dann aber findet kein vernünftiges Abruftraining statt, welches die stabilisierende Funktion des Testens ausmacht. Um eine unvernünftige Anwendung möglichst zu verhindern, wurden vor der Übung Empfehlungen gegeben, wie man im Einzelnen am besten vorgehen sollte. Vielleicht

würde ein Training oder zumindest eine gewisse Erfahrung mit den neuen Methoden die Qualität dieser Übungsmethode verbessern und so die Übungswirkung noch etwas erhöhen.

Selbsttesten erneutem Einprägen kognitiv und motivational überlegen

Wenn unter relativ ungünstigen statistischen Validitätsbedingungen immerhin eine Testmethode hypothesenkonform signifikant höhere Behaltenswerte nach sich zieht als Study only, so mag dies angenehm überraschen. Die Übungsmethode Selbsttesten war hier dem ausschließlichen Studieren via Landkarte signifikant überlegen. Der Behaltensvorteil des Selbsttestens von ca. 12 Prozent entspricht vom Ausmaß her einer mittleren Effektstärke. Der Computer ist im gegebenen Fall als schwer ersetzbares Lernmedium zu betrachten, da das praktizierte Selbsttesten auf Prozeduren beruht, die nur mit Computer praktikabel durchzuführen sind. In etlichen Computeranwendungen werden Möglichkeiten angeboten, mit der Maus ortsbezogen bestimmte Informationen anzufordern. Diese können dazu genutzt werden, sich selbst gezielt zu testen und zur Kontrolle die Information als Feedback abzurufen. Weitere Untersuchungen zur Bestätigung des hier gefundenen Effektes sind dringend notwendig, um die Stabilität und das Ausmaß des Behaltensvorteils besser abschätzen zu können. Da die hier verwandte Methode des Selbsttestens via CSA eine gezielte Informationsaufnahme keineswegs ausschließt - etwa: "Ich will jetzt ohne vorheriges Nachdenken direkt wissen, wie dieser Staat heißt und klicke deshalb in sein Staatsgebiet" - könnte der Vorteil gegenüber Study only auch dahin gehend interpretiert werden, es sei in bestimmten Lernstadien günstiger, dem Lerner selbst die Entscheidung zu überlassen, ob er sich gezielt informieren oder testen will.

Wenngleich die subjektive Bewertung der Gesamtübung unter Selbsttesten und Study only via Landkarte hoch vergleichbar ausfällt, sprechen die realisierten Nutzungszeiten beim Selbsttesten eindeutig dafür, der Lerner ziehe das Selbsttesten dem Studieren via Landkarte vor. D.h. die Studenten lernen lieber (relativ öfter bzw. deutlich länger) durch Selbsttesten als durch konventionelles Kartenstudium, wenn ihnen diese Wahlmöglichkeit angeboten wird. Dies lässt hoffen, Lerner würden freiwillig ein solches Übungsangebot auch eher nutzen. So gesehen bietet diese Form des Selbsttestens Lern - und Motivationsvorteile gegenüber einfachem Kartenlesen.

Aufwändige Testprozedur nicht besser als einfache Testvariante

Erneut hat sich gezeigt, dass die aufwändigere Testübungsvariante (MC-Flashcard mit response contingent-feedback) einer einfachen Testform (CSA mit KCR-Feedback) nicht zwingend überlegen ist, wenn letztere ein vernünftiges Üben zulässt. Es kommt vermutlich nicht allein darauf an, ein falsches Item möglichst oft, sondern möglichst gut einzuprägen. Unter der MC-Methode wurde zudem stets response contingent feedback gewährt. Wenngleich bisherigen Befunden nach RCF meistens keinen höheren Lernerfolg bewirkte als KCR (siehe Jacobs 2004), waren die theoretischen Erwartungen für den potentiellen Vorteil von RCF ziemlich günstig, weil die Verwechslungen ebenfalls zum Lehrziel gehören und nähere Klärung für 2 Items anboten, die es voneinander zu unterscheiden galt. Aber unter der einfachen Übungsvariante CSA hatte der Lerner offenbar genügend Zeit, diejenigen Items länger einzuprägen, die ihm problematisch erschienen. Bei durchschnittlich 2 realisierten Übungsdurchgängen unter CSA kann dies durchaus genügen bzw. nicht schlechter sein als Flashcard mit RCF-Feedback.

Ähnlich wie bei Jacobs (2006) gelang es auch hier nicht, den numerischen Behaltensvorteil der aufwändigen MC-Testvariante gegenüber Study Only statistisch zu sichern. Doch finden sich hier, wenn auch in schwächerem Ausmaß als bei Jacobs (2006), statistische Belege für eine etwas höhere Wertschätzung der aufwändigen MC-Testprozedur gegenüber Study Only via Land-

karte. Insofern deuten sich für diese aufwändige Übungsmethode zumindest gewisse motivationale Vorteile gegenüber einfachem Kartenlesen an.

Mögliche Einwände gegenüber dem durchgeführten Experiment

Die relativ umfangreiche Anzahl von 50 zu lernenden Items war relativ anspruchsvoll und könnte mit dazu beigetragen haben, dass etliche Probanden ziemlich überfordert waren. Möglicherweise war das Retentionsintervall von einer Woche eine sehr strenge Forderung für das langfristige Behalten, insbesondere deshalb, weil das erlernte Faktenwissen wenig Querverbindungen untereinander zulässt und die Übung insgesamt eher als massierte Übung zu betrachten ist, was beides schnelles Vergessen begünstigt. Ursprünglich hatte ich auch geplant, die experimentellen Übungen zeitlich deutlich später als die Lernaneignungsphase zu gestalten, was jedoch 3 Untersuchungstermine beansprucht hätte und ohne finanzielle Unterstützung organisatorisch nicht zu bewältigen war. Nach Rohrer & Pashler (2007) wäre ein Interstudierintervall (ISI) von ca. einem Tag besonders geeignet gewesen, um die beste Behaltensleistung nach einer Woche Retentionsintervall (RI) zu erzielen, weil dann das Verhältnis von ISI zu RI im optimalen Bereich zwischen .1 und .2 liegt. Zu bedenken gilt weiterhin, dass die eigentliche Experimentalphase etwas mehr als die Hälfte der gesamten Lernaneignung ausmachte sowie lediglich 12 Minuten beanspruchte, und riesige Treatmentunterschiede in dieser relativ kurzen Übungszeit völlig unrealistisch sind.

Persönlichkeit als entscheidender Faktor erfolgreichen Einprägens?

Die praktisch bedeutsamsten Behaltensunterschiede basieren nicht auf den Übungsmethoden, sondern auf den Personen. Die Personenstreuung im langfristigen Behalten fällt gewaltig aus. Ähnlich sieht die Situation auch bei solchen Studien aus, welche pädagogisch anspruchvollere Lehrziele prüften (etwa Kang et. al. (2007) oder McDaniel et al. (2007)). Es kann hier schwer abgeschätzt werden, in wie weit das Vorwissen dafür mit verantwortlich ist. Aber, wenn eine vernünftige Informationsbasis und angemessene Übungsbedingungen vorgegeben werden, erscheint es offensichtlich sehr bedeutsam, wie der einzelne Lerner die verfügbare Information enkodiert, im Langzeitgedächtnis ablegt bzw. aus diesem abruft. Das strategische Einprägevorgehen, das diesem zugrunde liegende Eigenbemühen sowie die dafür erforderliche Fähigkeit sind deutlich wichtiger als die Form, wie raffiniert die Informationen im einzelnen erfragt oder zum Einprägen angeboten werden. Die durch unterschiedliche Lernmotivation bedingte Personenvarianz könnte man durch leistungsabhängige Geldanreize etwas eingrenzen (Carpenter & Pashler (2007); siehe auch Jacobs, 2007). Ein alternativer pädagogischer Ansatz bestünde darin, geeignete Memorierungs- und Lernstrategien einzuüben, die im Erfolgsfall den Vorteil mit sich brächten, auf weitere Lehrziele bzw. Lehrinhalte zu transferieren. Dies hört sich sehr schön an, ist aber nicht einfach zu erreichen, weil auch hier die Personen über unterschiedliche Fähigkeiten verfügen, diese Strategien umzusetzen.

Wichtiger Hinweis:

Eine mittlerweile durchgeführte Replikation zum Vergleich zwischen study only und selbstkontrolliertem Testen mit Feedback konnte den hier gefundenen Vorteil für das Selbsttesten nicht bestätigen. Siehe:
Jacobs, B. (2008). 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

Literatur

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. European Journal of Cognitive Psychology, 19, 514-527.

Carpenter, S. K., Pashler, H., Vul, E. (2006). What types of learning are enhanced by a cued recall test. Psychonomic Bulletin & Review, 13 (5), 826 - 830.

Carpenter, S. K. & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. Psychonomic Bulletin & Review, 14 (3), 474-478

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

Clifton, K. S. (2005). The Testing effect: using retrieval practice in the classroom. Thesis submitted to Marshall University In partial fulfillment of the Requirements for the degree of Master of Arts Psychology http://www.marshall.edu/etd/masters/clifton-karen-2005-ma.pdf [19.10.2005]

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

Duchastel, P. C. & Nungester, R. J. (1984). Adjunct question effects with review Contemporary Educational Psychology 9 (2) 97-103

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.

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. (2006). 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. (2007). Geld und Noten als extrinsische Motivatoren zur Verbesserung kognitiver Leistungen. URN: **urn:nbn:de:bsz:291-psydok-9644** URL: **http://psydok.sulb.uni-saarland.de/volltexte/2007/964/**

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.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. Journal of Educational Psychology, 67, 259–266.

Lenth, R. V. (2006). Java Applets for Power and Sample Size [Computer software]. Retrieved *06,28,2007* from http://www.stat.uiowa.edu/~rlenth/Power

McDaniel, M. A., Anderson, J. L., Derbish, M.H., & Morrisette, N. (2007). Testing the testing effect in the classroom. European Journal of Cognitive Psychology, 19, 494-513.

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

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. Psychonomic Bulletin & Review, 14, 474-478.

Roediger, H. L. & Karpicke, J. D.(2006a). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychological Science, 17 (3) 249-255.

Roediger, H. L. & Karpicke, J. D. (2006b). The Power of Testing Memory. Basic Research and Implications for Educational Practice. Perspective on Psychological Science, 1 (3) 181-210.

Rohrer, D. & Pashler, H. (2007). Increasing Retention without Increasing Study Time. Current Directions in Psychological Science. 16 (4), 183-186.

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

Thompson, C. P., Wenger, S. K. & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. Journal of Experimental Psychology: Human Learning and Memory, 4, 210−221. zitiert nach Roediger & Karpicke (2006b).

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

---

created: Bernhard Jacobs, 1.10..2007, b.jacobs@mx.uni-saarland.de

---

Anhang

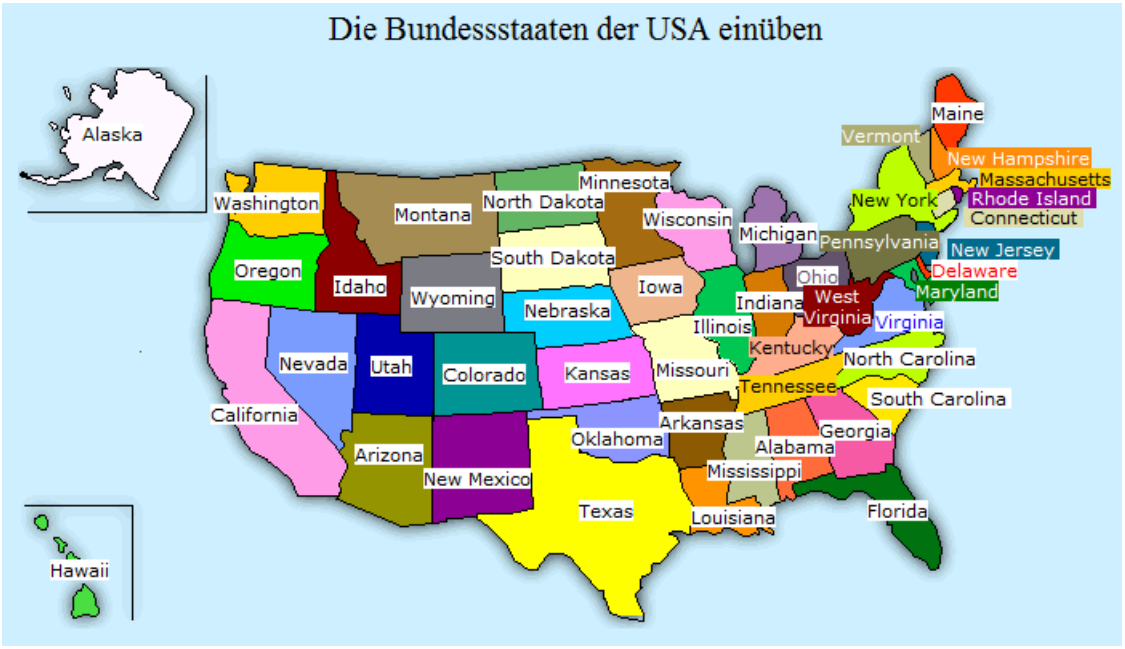
Flashcard-Methode: Multiple Choice Test mit KOR+KCR+ RCF Feedback+ Flashcard (MC): Fuzzy-Progammierung zur Verdeutlichung des Ablaufs:

```
procedure flashcard
{
 Bringe alle Items in eine zufällige Reihenfolge
 wiederhole
  {
    Teste das erste Item
    Gebe Feedback KOR+KCR  //= richtig/falsch sowie die Präsentation der korrekten Antwort
    wenn (erstes Item = richtig)
      {
        entferne das Item aus der Itemliste
      }
    wenn (erstes Item = falsch)
      {
        Gebe Feedback RCF
        Füge das Item an das Ende der Itemliste an
      }
   }
  bis die Itemliste kein Item mehr enthält
}

Zeige 30 sec alle falschen und verwechselten Staatsnamen in den entsprechenden Staatsgebieten
wiederhole procedure flashcard bzw. die Gesamtprozedur bis die Zeit abgelaufen ist.
```

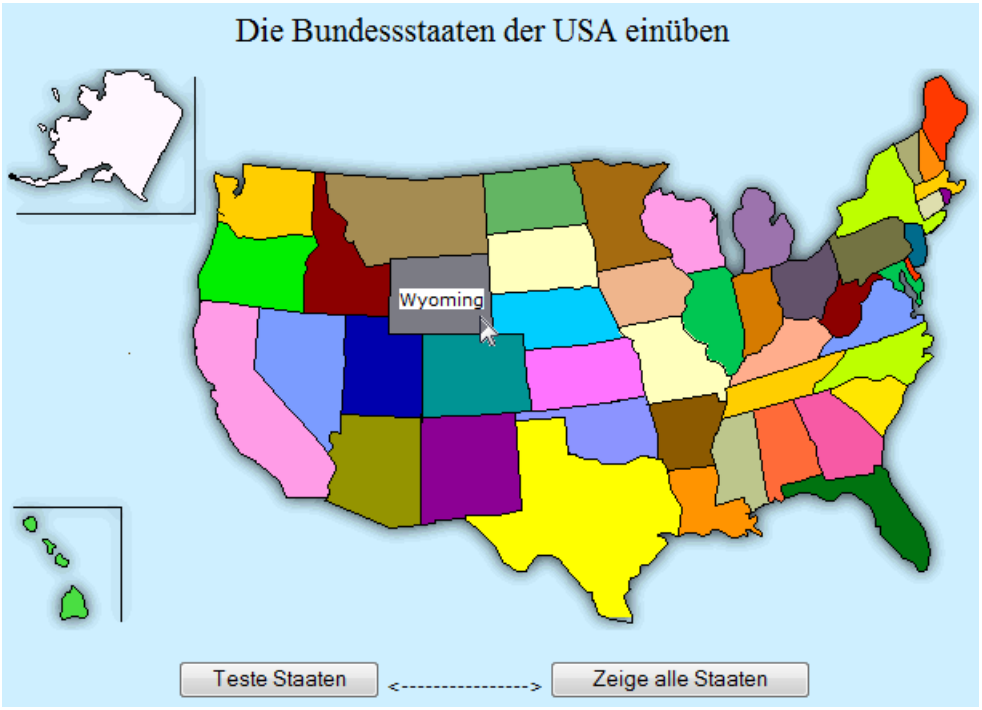
Bildschirmkopien der experimentellen Bedingungen

Erneutes Studieren mit der Landkarte: Study only (SO):



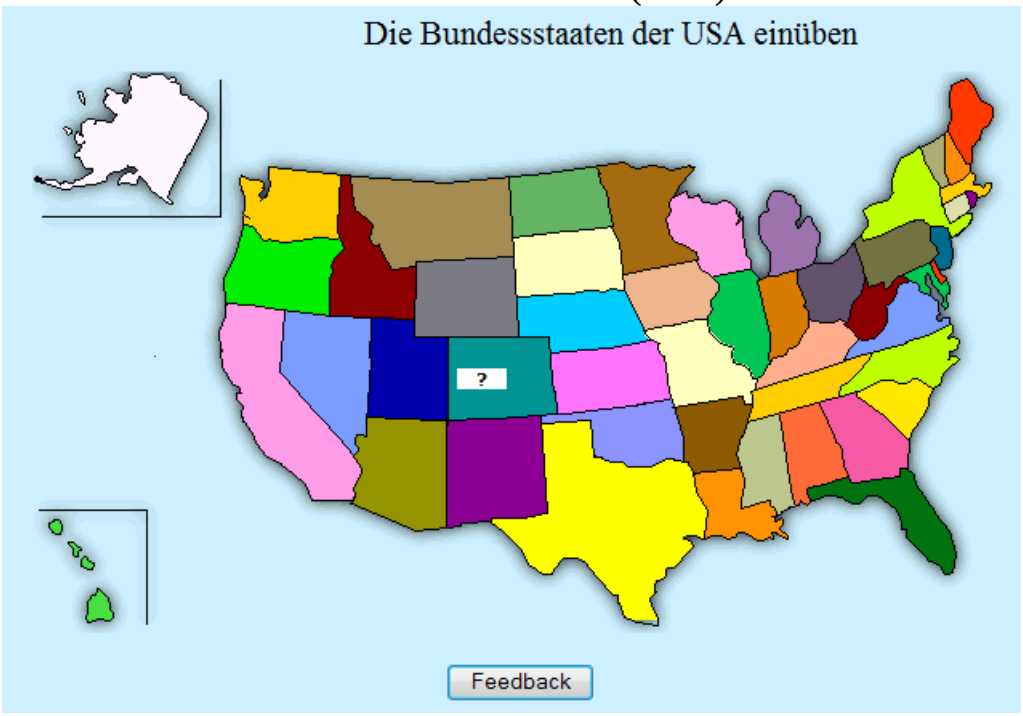
Der Lerner sieht nur diese Karte sowie die Anweisung
Wenden Sie eine sinnvolle Strategie an, sich möglichst viele Staaten einzuprägen

Selbsttesten mit CSA (Selbsttesten)



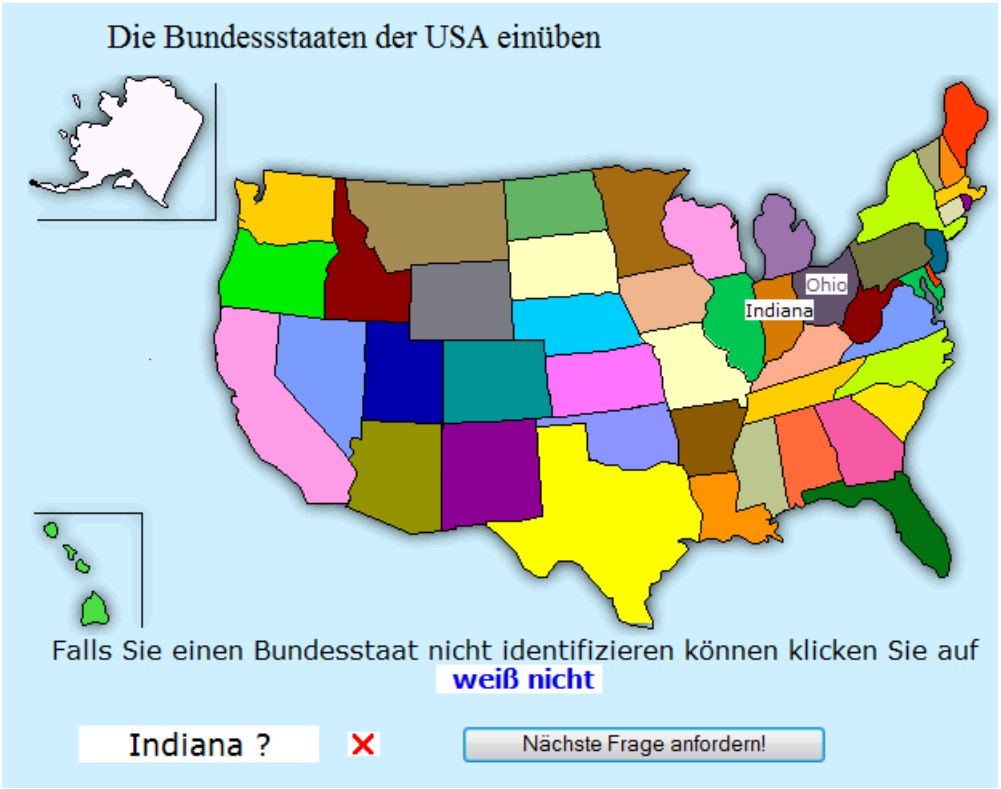
Der Lerner soll ein beliebiges Staatsgebiet auswählen und versuchen, sich an den Namen zu erinnern. Durch Anklicken des Gebietes von Wyoming sieht er hier den entsprechenden Namen. Beim Verlassen des Mauscursors aus dem Gebiet verschwindet der Name wieder und der Lerner kann das nächste Staatsgebiet auswählen. Zusätzlich hat der Lerner die Option, jederzeit in einen SO-Modus (siehe oben) zu wechseln und auf einen Blick alle Staatsnamen in den entsprechenden Gebieten zu sehen

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 auf Button Feedback oder durch Tippen auf die Leertaste. Daraufhin erscheint an der Stelle des Fragezeichens hier der Staatsname Colorado (=KCR)

Multiple Choice Test mit KOR+KCR+ RCF Feedback + Flashcard: (MC):



Der Lerner sollte mit der Maus auf Indiana klicken, hat aber Ohio angewählt. Deshalb erhält er symbolisch das Feedback falsch (KOR). Nun sieht er den Namen Indiana im zutreffenden Gebiet von Indiana (KCR). Zusätzlich wird ihm verdeutlicht, wo Ohio liegt (RCF). Die spezielle Flashcardmethode fügt falsche Items ans Ende der Übungsliste und testet diese solange, bis ein korrektes Ergebnis erzielt wird.