

**Bernhard Jacobs**, Fachrichtung Bildungswissenschaften der Universität des Saarlandes.  
 Email: b.jacobs@mx.uni-saarland.de  
 Version: 16.1.2012

## Auf der vergeblichen Suche nach dem Testeffekt - Studieren oder Testen mit Feedback beim Vokabellernen.

### Abstract

Ziel der Studie war die Überprüfung einer bisher häufig bestätigten Hypothese, beim Einüben von Vokabeln bewirke Testen mit Feedback einen höheren langfristigen Lernerfolg als wiederholtes Studieren. Studierende wurden zufällig auf drei Übungsbedingungen zugeteilt: erneutes Studieren, klassisches Short-Answer-Testen mit unmittelbarem Feedback sowie einer weiteren Testmethode, welche die schriftliche Eingabe durch die gedankliche Antwort ersetzte und im Anschluss als Rückmeldung die korrekte Antwort gewährte. Die Probanden nahmen via Internet an der Übung teil und konnten ihre Übungszeit selbst bestimmen. Der 5 Tage später anberaumte Behaltens- und Transfertest sowie ein noch später eingesetzter free recall test ergaben keine Behaltensvorteile für die Testvarianten, sondern deuteten eher Nachteile beim Transfertest an. Ebenso ließen sich zwischen den Übungsmethoden keine Unterschiede hinsichtlich Höhe und der Genauigkeit der subjektiven Leistungseinschätzungen feststellen. Die Untersuchung kann somit weder den strengen Testeffekt, noch metakognitive Vorteile des Testens im Bezug auf die Einschätzung des eigenen Wissens bestätigen. Die Divergenzen zu den bisherigen Ergebnissen werden zum Teil im methodischen Vorgehen vermutet, das zugunsten praxisnaher Lernbedingungen auf einige Reglementierungen der Laborforschung verzichtete. Das Fehlen der Treatmentunterschiede wird überwiegend damit erklärt, alle eingesetzten Methoden gewährten vergleichbar gute Chancen zum Einprägen und Behalten und die Behaltensergebnisse würden im Wesentlichen von der Fähigkeit und der Lernbereitschaft der Person abhängen, die jeweilige Übungsgrundlage angemessen zu nutzen.

**Schlagworte:** Üben, Testen, Feedback, Aufgabentypen, Vokabellernen

### Einleitung

Tests dienen nicht nur der Diagnose des Wissens, sondern auch als Übungsgelegenheit zur Stärkung des zuvor Gelernten. Wer nach einer Instruktionsphase einen entsprechenden Test bearbeitet, erzielt zu einem späteren Zeitpunkt bessere Behaltenswerte als der, welcher keinen Test bearbeitete (= lenient criterion test effect nach Cranney et al., 2009; siehe z.B. Hamaker 1986, Jacobs 2006). Dieser einfache Testeffekt entfaltet sich insbesondere bei hohen Erfolgsquoten im Übungstest. Die Lernwirksamkeit steigt in der Regel an, wenn sich der Testung ein Feedback anschließt, das die korrekte Lösung umfasst, weil so mögliche Fehler korrigiert werden können. Mit wachsender Fehleranzahl im Übungstest erhöht sich die Dringlichkeit, die korrekte Antwort rückzumelden. Während die Lernwirkung des Testens mit Feedback unstrittig ist, stellt sich die Frage, ob nicht alternative Übungsmethoden eine vergleichbare Lernwirkung hinterlassen. In diesem Zusammenhang erweist sich die gezielte erneute Präsentation der lehrzielrelevanten Informationen als naheliegende Methode. In letzter Stringenz steht hier ein Vergleich zwischen der Testung mit Feedback und der direkten Darbietung der Frage einschließlich der korrekten Antwort an. Beim Vokabellernen würden statt einer Testung die Vokabelpaare erneut zum Studieren angeboten oder bei Mathematikaufgaben direkt eine Musterlösung vorgelegt, welche Fragestellung, korrekte Antwort und das beim Testen ansonsten noch gewährte Feedback anbietet.

Mittlerweile liegen relativ viele Laborstudien vor, welche die Überlegenheit des Testens mit und teilweise sogar ohne Feedback gegenüber der erneuten Präsentation der Lerngrundlage bei unterschiedlichen Lernarten nachgewiesen haben [=strict criterion testing effect nach Cranney et al. 2009]. Der Nachweis eines solch strengen Testeffektes gelang insbesondere

ziemlich konsistent beim Erlernen einfachen Faktenwissens, das vornehmlich mit Aufgabentypen erfasst wurde, die eine eigene Antwort verlangen, sogenannten Constructed-response-Aufgaben wie short answer [kurze Freiantwortaufgabe] oder free recall [freie Wiedergabe]. Dazu gehört das Einüben von Vokabeln oder sonstigen Paarassoziationen: z.B. Carrier & Pashler, 1992; Cull, 2000; Carpenter & DeLosh, 2005; Carpenter, Pashler & Vul 2006; Jacobs, 2006, Karpicke, 2007; Toppino & Cohen, 2009, Wei-chieh Fang, 2010, Pyc & Rawson, 2010, Vilhelmsson, 2011). Flashcardvarianten, welche das Testen gegenüber dem Studieren favorisierten, ergaben bessere Ergebnisse als Flashcardmethoden, die mehr Studieritems statt Testitems vorsahen. (Karpicke & Roediger, 2008; Schmidmaier et al. 2011). Die klaren Befunde liefern offenbar hinreichende wissenschaftliche Belege dafür, beim Vokabellernen oder sonstigem trivialen, reproduzierenden Faktenwissen statt erneutem Studieren verstärkt Testungen einzuführen, um das langfristige Behalten zu stärken. Streng genommen setzt diese Empfehlung allerdings den Zusatz voraus, "wenn in der pädagogischen Praxis auch so gelernt wird, wie dies in den Laborexperimenten normalerweise geschieht." Durch nachfolgende Studie sollte überprüft werden, ob sich die Ergebnisse auch auf Situationen übertragen lassen, die dem Lernen unter üblichen Alltagsbedingungen eher entsprechen. Damit sollte ein Aspekt der externen Validität ins Visier genommen werden.

## Laborstudien und natürliches Lernen

Viele Experimente sind eher am Nachweis eines wissenschaftlichen Effektes und weniger an einer praktischen Verwertbarkeit der Ergebnisse interessiert, was mitunter dazu führt, dass die Forscher ihren Probanden Lernbedingungen zumuten, die diese in der Praxis niemals anwenden würden. Ganz im Sinne der "Maximierung der UV-varianz" werden die Lernenden massiv mit Testungen konfrontiert, die man zum einen in der Praxis nicht realisieren kann, etwa die Aufforderung an die Übenden, die wesentlichen Aussagen eines einmal gelesenen Textes dreimal hintereinander mit eigenen Worten wieder zu geben (z.B. Roediger & Karpicke, 2006) oder die sich im Schulsystem langfristig nicht durchsetzen lassen, etwa eine Testung unmittelbar vor dem Unterricht, unmittelbar danach sowie 2 Tage vor dem Examen. (McDaniel et al. 2011). Was für die experimentelle Forschung Vorteile bietet, etwa die Verwendung eines Repeated measurement Designs findet selten Entsprechung im normalen Übungsvorgehen von Schüler und Studenten. Viele Studien basieren auf Wiederholungsexperimenten. Hierbei erhalten die Probanden bei einem Teil der Aufgaben die Studiermethode, bei einem anderen Teil die Testmethode. Dies könnte Kontraste hervorrufen, welche beispielsweise die Studieraufgaben als langweilig und uninteressant und die Testaufgaben als spannend und herausfordernd auffassen ließen. Jacobs (2006) fand in einem solchen Wiederholungsdesign z.B. deutliche Hinweise auf eine Präferenz von Testmethoden gegenüber einfachem Studieren, konnte aber in einem unabhängigen Gruppenplan (Jacobs 2007) solche Präferenzunterschiede nicht aufdecken.

## Die Bedeutung der Lernzeit für den Vergleich von Übungsmethoden

Da Lernerfolge von Übungsmethoden nur bei vergleichbarer Lernzeit vernünftig interpretierbar sind, versuchten etliche Forscher die Lernzeiten beider Übungsmethoden konstant zu halten. So legt man etwa für das Testen eines Vokabelpaares die Testzeit auf 6 Sekunden und die Feedbackzeit auf 4 Sekunden fest, während bei der Studiermethode dann das Vokabelpaar 10 Sekunden auf dem Bildschirm verbleibt. Hierbei wird weder Rücksicht auf das unterschiedliche Memorierungsvermögen der Probanden oder die jeweilige Schwierigkeit der Aufgabe genommen, noch der unterschiedliche Zeitbedarf der Übungsmethoden in Betracht gezogen. Schriftliches Testen mit anschließendem Feedback erfordert aber deutlich mehr Zeit als das Studieren eines Vokabelpaares. Manche Forscher (z.B.: Carpenter & DeLosh 2005, Wei-

chieh 2010) kamen vermutlich deshalb auf die Idee, auch beim Studieren eines Vokabelpaares die Probanden damit zu beschäftigen, eine Ihnen vor Augen liegende Vokabel schriftlich in ein Antwortfeld einzugeben. Man kann darüber streiten, ob diese Maßnahme eine effiziente Form der Enkodierung darstellt. Konstante Übungszeiten für alle Items erweisen sich insofern als problematisch, als gute Lerner weniger Zeit als schwache Lerner und leichte Vokabeln weniger Übungszeit als schwierige Vokabeln erfordern (z.B. Jacobs 2009, S.15). In vielen Studien wird auf die Mitteilung von Übungszeiten schlichtweg verzichtet. Man darf annehmen, in diesen Fällen sei der Vorteil des Testens vermutlich mit einer höheren Übungszeit konfundiert gewesen. Vilhelmsson (2011) macht zumindest gewisse Angaben zur Lernzeit beider Übungsmethoden und gibt 5 Sekunden Lernzeit für das Studieren der Vokabeln, aber bis zu 30 Sekunden für die Testung an. Etliche Programmvarianten setzen die Üben massiv unter Lernstress. Hierbei laufen die Übungsanforderungen ohne erkennbare Selbststeuerung des Üben automatisch ab, verlangen vom Lerner die volle Konzentration und planen keine hinreichende Zeit zur Reflexion oder Elaboration ein.

Nur bei freier Lernzeit hat der Übende die Chance, seine Übungszeit den verschiedenen Anforderungen angemessen anzupassen, wenngleich diese Freiheit die vernünftige selbst gesteuerte Adaptation natürlich nicht automatisch garantiert. Aber solange keine überzeugenden Belege für eine Begrenzung der Übungszeit pro Item vorgelegt werden sowie Studierende sich in ihrer normalen Lernumgebung auch nicht dem Diktat einer Zeitbeschränkung unterwerfen, erscheint es ratsam, sie selbst entscheiden zu lassen. Jacobs (2006, 2007, 2008, 2009) versuchte das Zeitproblem dadurch zu lösen, die Gesamtübungszeit für das Studieren und Testen konstant zu halten. Es blieb den Probanden überlassen, wie lange sie sich mit einer Aufgabe befassen wollten, aber die Übung wurde zum festgelegten Zeitpunkt abgebrochen. Hierbei konnte Jacobs (2006) z.B. feststellen, dass bei konstanter Gesamtübungszeit mehr als doppelt so viele Aufgaben studiert wie durch einen Short Answer Aufgabentyp getestet worden waren. Gelegentlich zeigte sich bei diesem Verfahren die erhoffte Überlegenheit des Testens gegenüber dem Studieren. In der Gesamtbetrachtung aller Ergebnisse ließ sich jedoch kein deutlicher Vorteil für das Testen ausmachen. Auch dieses Zeitverfahren mag gewisse Nachteile haben, weil so bei einer Methode zwingend mehr Lerndurchgänge als bei der anderen Methode vorliegen. Bei freier Aufgabenübungszeit sind gleiche Lernzeiten von Studieren und Testen bei gleicher Anzahl von Lerndurchgängen allerdings nicht realisierbar.

## Ziel der Studie

Ziel dieser Studie ist es, den strengen Testeffekt, nämlich die höhere langfristige Behaltenswirksamkeit des Testens mit Feedback gegenüber dem erneuten Studieren bei einfachem Faktenwissen zu überprüfen und hierbei die Übungszeiten jeder Aufgabe der freien Entscheidung des Einzelnen zu überlassen, aber dennoch eine gewisse Vergleichbarkeit der Übungszeiten anzustreben. Da die bisher durchgeführten Studien den strengen Testeffekt bei einfachem Faktenwissen mehrfach bestätigt hatten, gilt auch hier die Hypothese, Testen mit Feedback fördere das langfristige Behalten mehr als erneutes Studieren.

Um eine dem Studieren der Vokabeln annähernd zeitgleiche Testmethode zu verwenden, wurde neben einer klassischen Testung mit Feedback eine Testvariante eingesetzt, die ein explizites Schreiben durch die gedankliche Antwort ersetzte. Diese von Jacobs (2006) als covert short answer bezeichnete Testmethode kam auch in der Studie von Carpenter, Pashler & Vul (2006) zum Einsatz und wurde dort covert retrieval genannt. Weiterhin sollte der bereits von Carpenter, Pashler & Vul (2006) sowie Wei-chieh (2010) nachgewiesene Vorteil des Testens gegenüber dem erneuten Studieren im Hinblick auf einen minimalen Transfer beim Paarassoziationslernen untersucht werden, jetzt aber unter veränderten Untersuchungs-

bedingungen. Meistens kamen in der Test- und Feedbackforschung identische Aufgaben in Übung und Behaltenstest zum Einsatz, was den Nachweis eines Lernerfolgs auf eine reine Memorierungsleistung einengt. Auch hier werden beim Transfertest dieselben Vokabeln herangezogen wie in der Übungsphase, allerdings in einer veränderten Fragerichtung. In der Übungsphase wird beim Testen stets die Fremdsprachenvokabel vorgegeben und nach der deutschen Vokabel gefragt. Im Transfertest dann aber die deutsche Vokabel vorgegeben und die Fremdsprachenvokabel verlangt. Die veränderte Fragestellung überbrückt gegenüber der Übungsfrage natürlich nur eine sehr geringe Transferdistanz, dürfte für manche Vokabeln allerdings keine ganz triviale Anforderung darstellen. Sollte Testen über reine Memorierung hinaus wirken, dann müssten die Testmethoden, so sie im identischen Behaltenstest besser abschneiden als die Studiermethode, auch beim Transfertest höhere Leistungen erbringen.

## Die verwendeten Vokabeln

Wie in der Laborforschung üblich, sollte das einzuübende Lernmaterial unbekannt sein, weswegen die in der Memoryforschung so beliebte ostafrikanische Fremdsprache Swahili, nach <http://www.swahili.de/> auch Kiswahili, Suaheli oder Kisuaheli genannt, zum Einsatz kam. Der Verfasser dieser Arbeit suchte mehr oder weniger willkürlich insgesamt 20 konkrete wie abstrakte Worte aus unterschiedlichen Gebieten aus (siehe Anhang A). Alle Swahili-Worte sind Substantive mit wohlklingender Aussprache, umfassen 2 oder 3 Silben und stellen keine hohen Anforderungen an die Rechtschreibung (z.B.: hewa, jambo, ramani, subira).

## Die experimentellen Übungsmethoden

Neben der Studiermethode kamen zwei Testvarianten zum Einsatz. In Tabelle 1 werden die Übungsmethoden etwas genauer besprochen. Bei der klassischen Testvariante musste der Proband auf Vorgabe der Fremdsprachenvokabel die deutsche Vokabel in ein Antwortfeld schreiben. Diese Form der Testung basiert auf dem Short Answer Aufgabentyp (deutsch: kurze Freiantwortaufgabe).

**Tabelle 1: Beschreibung der Übungsvarianten und Link zur entsprechenden Übungsmethode**

Übungsvarianten (Bildschirmkopien siehe Anhang B)	Beispiel
<b>SO Study only</b> , auch Studieren genannt: Ein Vokabelpaar wird auf dem Bildschirm präsentiert und der Proband kann selbst entscheiden, wie lange er sich dieses Vokabelpaar einprägen will. Durch Anklicken auf den Button "nächste Vokabel" wird das nächste Vokabelpaar angefordert.	<a href="#"><u>Study only (So)</u></a>
<b>CSA Covert Short Answer Testen mit KCR-Feedback:</b> Auf dem Bildschirm erscheint die Fremdsprachenvokabel. Der Proband sollte sich an die deutsche Vokabel erinnern und diese leise aussprechen. Durch Mausklick auf den Button "Korrekte Antwort" oder Klick auf die Leertaste erscheint dann die deutsche Vokabel. Ein Klick auf den inzwischen veränderten Button "nächste Vokabel anfordern" präsentiert die nächste Fremdsprachenvokabel.	<a href="#"><u>Covert Short Answer (CSA)</u></a>
<b>SA Short Answer Testen mit KCR-Feedback:</b> Auf dem Bildschirm erscheint die Fremdsprachenvokabel und ein Antwortfeld. Der Proband soll die deutsche Vokabel eingeben und dann die Taste "Aufgabe bestätigen" anklicken. Unmittelbar danach erhält er symbolisch die Rückmeldung "richtig/falsch" und zusätzlich im Feedbackfeld die korrekte deutsche Vokabel mitgeteilt. Zu diesem Zeitpunkt sind also die Fremdsprachenvokabel, die Antwort des Probanden und die korrekte Antwort sichtbar. Ein Klick auf den Button "Nächste Vokabel anfordern" zeigt die nächste Fremdsprachenvokabel auf dem Bildschirm.	<a href="#"><u>Short Answer (SA)</u></a>

Bei der zweiten Testvariante lag dieselbe Aufgabenstellung zugrunde, nur sollte der Proband die deutsche Vokabel lediglich erinnern bzw. leicht aussprechen. Dadurch entfällt die

Schreibarbeit, was Zeit einspart. In beiden Testvarianten folgte auf Anforderung des Probanden als Rückmeldung die korrekte Antwort, in etlichen Feedbackklassifikationen knowledge of correct response oder knowledge of correct result genannt (z.B. Jacobs, 2002).

Die SA-Testung erwartete zwar eine Antwort, bestand aber nicht darauf, so dass das Antwortfeld auch leer bleiben konnte. Die angeregte gedankliche oder leise ausgesprochene Antwort unter CSA entzieht sich jeder Kontrolle, weswegen die Korrektheit der Antwort weder erfasst noch rückgemeldet werden konnte. Die zutreffende Antwort wurde aber unter beiden Testvarianten zwingend rückgemeldet und blieb solange auf dem Bildschirm, bis der Proband die nächste Aufgabe anforderte.

## Die abhängigen Variablen

Bei allen Bedingungen wurden die **Bearbeitungszeiten** für die beiden Übungsphasen gemessen, unter SA zusätzlich die Anzahl der korrekten Lösungen pro Übungsdurchgang. Aus Gründen einer besseren Vergleichbarkeit über diverse Studien hinweg wird statt der Anzahl der korrekten Lösungen stets der Prozentsatz der korrekten Lösungen kommuniziert. Als wichtige abhängige Variablen gelten die via Onlineerhebung erfassten Behaltenstests, zunächst in der eingeübten Reihenfolge (identischer **Behaltenstest**) und anschließend in der umgekehrten Reihenfolge (**Transfertest**). Tabelle 2 stellt die Fragerichtung für die Übungs- und Testphasen gegenüber.

**Tabelle 2: Fragerichtung bei der Übung und den Behaltenstests**

	<b>cue</b> Vorgabe	<b>target</b> erwartete Antwort
Übungstest:	Fremdsprache	Deutsch
<b>Behaltenstest:</b>	Fremdsprache	Deutsch
<b>Transfertest:</b>	Deutsch	Fremdsprache

Um die Testmotivation bei den Behaltenstests anzuregen, war den Studierenden eine unmittelbare Rückmeldung des Prozentsatzes ihrer korrekten Lösungen versprochen worden. In einem späteren **free recall test** forderte der Seminarleiter die Studierenden auf, alle Vokabelpaare aufzuschreiben, an die sie sich noch erinnern konnten. Zudem sollten sie auch solche Vokabeln aufschreiben, deren Übersetzungskorrelat sie nicht mehr erinnerten. Aus diesen Angaben wurden 2 free recall tests gebildet.

- Free recall exact = Anzahl der zutreffenden Vokabelpaare
- Free recall total = Anzahl aller erinnerten Vokabeln, unabhängig von einer realisierten oder korrekten Zuordnung.

Während alle Übungs- und Onlinetestungen korrekte Schreibweise der erfragten Vokabeln erforderten, wurden im free recall test geringfügige Schreibfehler bei der Fremdsprachenvokabel toleriert, z.B.: pata statt bata, marziwa statt maziwa. Da beide Free Recall Variablen  $r=.94$  miteinander korrelieren, wurde nur der free recall exact als Messvariable verwendet, der im Folgenden als free recall test bezeichnet wird.

Neben Leistungsdaten wurden subjektive Schätzungen zum Behaltenstest erfasst, und zwar unmittelbar vor und nach dem Behaltenstest:

- Judgement of learning (Jol): "Schätzen Sie ein, wie viel Prozent der Vokabeln Sie vermutlich richtig lösen werden?"
- Judgement of performance (Jop): "Schätzen Sie ein, wie viel Prozent der Vokabeln Sie vermutlich richtig gelöst haben?"

## Probanden und Untersuchungsvorgehen

An der Studie nahmen Studierende des Lehramts aus 4 Seminaren teil, die der Verfasser im WS 11/12 leitete. Das Durchschnittsalter betrug 23 Jahre. Ca. 2/3 der Probanden waren Frauen. Den Studierenden wurde im Vorfeld mitgeteilt, sie sollten an einer Onlineerhebung teilnehmen, die den Zweck verfolgte, Anschauungsmaterial und Datengrundlage für etliche diagnostische Berechnungen sowie die Evaluation von Maßnahmen zu liefern. Etwa 90% aller SeminarteilnehmerInnen beteiligten sich an der Untersuchung. In einer kurzen Vorerhebung, die den Studierenden als Versuch deklariert wurde, die Funktionsfähigkeit des Onlineablaufs zunächst zu überprüfen, wurde der Abiturnotendurchschnitt erfragt, um ihn als Parallelisierungsvariable für den Versuch zu nutzen. 85% machten hierzu entsprechende Angaben.

85% aller Probanden wurden zunächst nach Abiturnotendurchschnitt in Tripel eingeteilt und innerhalb jedes Tripels nach Zufall den experimentellen Variablen zugewiesen. Die verbliebenen 15 % wurden nach dem klassischen Randomisierungsverfahren zugeteilt. Mithin liegt der Studie der in Tabelle 3 dargestellte, klassische experimentelle Versuchsplan ohne Vortest zugrunde. Zudem garantiert er durch die weitgehende Parallelisierung des Abiturnotendurchschnitts eine vergleichbare schulische Leistungsfähigkeit der Gruppen.

### Tabelle 3: Versuchsplanformalisierung

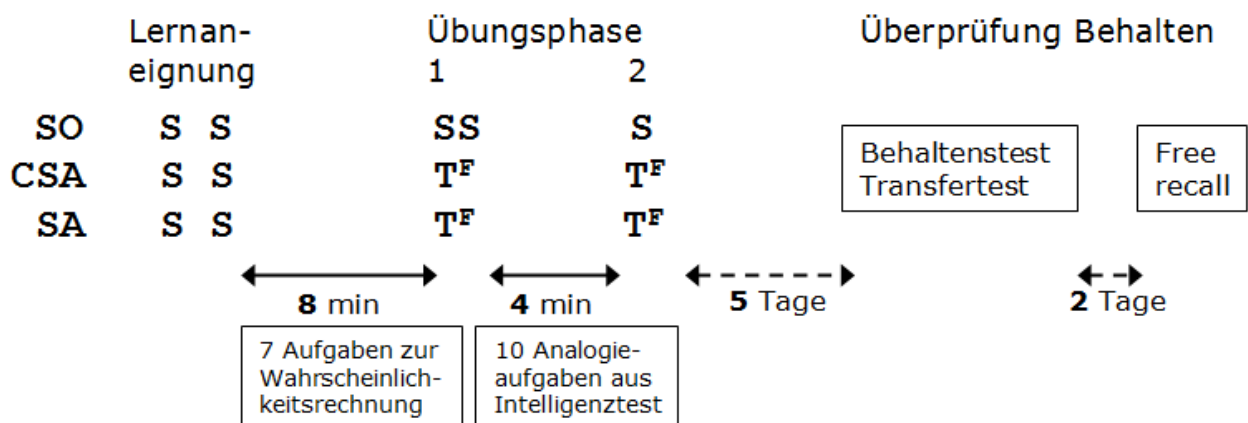
R	SO : erneutes Studieren	O
R	CSA: Testen mit Feedback	O
R	SA : Testen mit Feedback	O

Eine nachträgliche Überprüfung ergab vergleichbare Abiturnoten der experimentellen Gruppen in Deutsch- Mathematik sowie im Durchschnitt. Allerdings erwies sich diese Parallelisierung im Nachhinein als unnötig, da der Zusammenhang zwischen den Behaltenstests und allen Abiturnoten insignifikant ausfiel.

Die Organisation der Studie wurde über Email und Internetbrowser abgewickelt. Hierbei gewährte der Versuchsleiter den Probanden einen gewissen zeitlichen Rahmen, die Übung und die späteren Behaltenstests durchzuführen. Entsprechend der experimentellen Zuordnung bekamen die Probanden zunächst eine Email zugesandt, welche URL, Username und Passwort für die betreffende Übung enthielt sowie den Zeitrahmen für die Übungsrealisation absteckte. Drei Tage nach dem letzt möglichen Übungstermin forderte der Versuchsleiter die Probanden via Email auf, die zuvor nicht angekündigten Behaltenstests via Internetbrowser anzusteuern. Die Onlinebehaltenstests wurden frühestens 3 und höchstens 7 Tage nach der Übung in Angriff genommen. Das nach arithmetischem Mittel wie Median erfasste, durchschnittliche Behaltensintervall betrug fünf Tage. Ca. ein bis drei Tage nach den Onlinetests bearbeiteten die Probanden den free recall test im Seminar.

Die Abbildung 1 verdeutlicht den Ablauf der Untersuchung. Die Lernaneignungsphase gestaltete sich für alle Bedingungen identisch. Zu Beginn erhielten die Probanden die Information, sie bekämen 20 Vokabeln zum Einprägen präsentiert und diese Vokabeln würden später auch getestet werden. Zunächst sahen die Probanden für 2 Minuten eine Liste mit den 20 einzuübenden Vokabelpaaren. Diese Vokabeldarbietung in Listenform wird hier als ein Studierdurchgang S gewertet und gewährte eine durchschnittliche Bearbeitungszeit von 6 Sekunden pro Vokabelpaar. Der zweite Lernaneignungsdurchgang S war voll automatisiert. Jedes Vokabelpaar erschien einzeln für jeweils fünf Sekunden an der gleichen Stelle auf dem Bildschirm. Danach schloss sich eine Pause von einer Sekunde an, dem dann das nächste Vokabelpaar folgte. Die Lernaneignungsphase 1 dürfte eine den Probanden bekannte Lerngrundlage analog eines Vokabelheftes repräsentieren, während die Zeitbegrenzung das übliche Lernverhalten stört. Die zweite Lernaneignung hat artifiziellen Laborcharakter, dessen Funktion vorwiegend darin besteht, mehr oder weniger sicher zu stellen, dass der Übende auch jede Vokabel zur Kenntnis nimmt.

**Abbildung 1: Ablauf der Untersuchung**



Nach der Lernaneignung bearbeiteten die Studierenden 7 Aufgaben aus dem Fragebogen zur Wahrscheinlichkeitstheorie (Nachtigall & Wolf, 2001), jede experimentelle Gruppe unter einer unterschiedlichen Instruktionsbedingung, was hier nicht von Interesse ist. Da für diese Aufgaben keinerlei Zeitbeschränkung vorgesehen war, beziehen sich die 8 Minuten auf die durchschnittliche Bearbeitungszeit aller Probanden.

Die Übungsphasen legten lediglich fest, wie viele Lerndurchgänge zu absolvieren waren, wobei die Aufgabenbearbeitung vollständig der Eigensteuerung des Einzelnen unterlag. Vor jedem Übungsdurchgang erhielt der Proband nähere Erklärungen, was ihn erwartet und wie er vorgehen sollte. Die erste Übungsphase unter der Bedingung Studieren umfasste 2 Übungsdurchgänge, also 40 Vokabeldarbietungen, während die Testvarianten jeweils nur einen Durchgang absolvierten, der bei jeder Aufgabe aus Testen und unmittelbarem Feedback bestand. Der ersten Übungsphase schloss sich ein auf 3 Minuten beschränkter Analogietest an, der nach der Auswertung die korrekten Lösungen zur Ansicht anbot und dessen Gesamtbearbeitungszeit auf ca. 4 Minuten eingeschätzt wurde. Dann folgte die für alle Bedingungen auf jeweils einen Übungsdurchgang beschränkte Übungsphase 2. Für alle Vokabelpräsentationen der Lernaneignungs-, Übungs- und Behaltensphase galt: Die 20 Vokabeln wurden für jeden Probanden stets in zufälliger Reihenfolge dargeboten.

## Ergebnisse

### SA-Testen mit Feedback in der experimentellen Übungsbedingung

Um eine möglichst solide Einschätzung der Übungsdurchführung zu erzielen, gehen in die nachfolgende Analyse die Daten aller Studierenden ein, auch solcher Studierenden, die zu den späteren Behaltenstests nicht erschienen waren, zumal die Ergebnisse sehr hoch vergleichbar ausfallen, wenn man die 7 bei den Behaltenstests säumigen Probanden ausschließt. Die in Tabelle 4 dargestellten Daten unter der SA-Bedingung geben Hinweise auf die Reliabilität des Vokabeltests, das Behalten nach der Lernaneignungsphase und die Leistungsveränderungen durch Testen mit Feedback.

**Tabelle 4: Vokabeltest für Bedingung SA-Testen mit KCR-Feedback.** (N jeweils 34)

	Übungs- phase	M	s	$\alpha$	r	t	d
% korrekt	1	49.4	25.1	0.85			
					.95	-6.6	.40
% korrekt	2	59.7	28.6	0.91			
Zeit in sec 1	1	185	75				
					.74	4.8	.62
Zeit in sec 2	2	144	55				
Korrelation zwischen % korrekt und Bearbeitungszeit							
Übungsphase 1	r =	.34*					
Übungsphase 2	r =	.54**					

Die Testwerte im ersten Durchgang der experimentellen Bedingung SA von ca. 50% korrekter Antworten liefern eine tragfähige Schätzung für die Behaltenswirkung durch die Lernaneignungsphase und zwar für alle am Versuch beteiligten Probanden, weil diese ja nach Zufall den experimentellen Bedingungen zugewiesen wurden. Eventuell fällt die Schätzung etwas zu hoch aus, da den Vokabeln unmittelbares Feedback folgte, wodurch bestimmte Lernhinweise für die noch zu testenden Vokabeln des ersten Durchgangs gegeben wurden. In der zweiten Übungsphase erzielten die Probanden der SA-Bedingung ca. 10% höhere Behaltenswerte. Der Unterschied fällt nach t-Test für abhängige Stichproben - auch wegen der sehr hohen Retestkorrelation - hoch signifikant aus, entspricht aber nur einer Effektstärke von  $d = .40$ . Diese Behaltensverbesserung geht höchstwahrscheinlich auf das KCR-Feedback zurück, da bei einer reinen Testung im zweiten Übungsdurchgang höchstens das Ausgangsniveau des ersten Durchgangs erreicht worden wäre (siehe z.B. Karpicke & Roediger, 2007). Die Testbearbeitungszeit vollzieht sich beim zweiten Übungsdurchgang signifikant schneller und reduziert sich durchschnittlich von ca. 9 Sekunden pro Item auf ca. 7 Sekunden pro Item. Die Streuungen der Lernzeiten belegen aber große interindividuelle Unterschiede, die zumindest teilweise auf ein unterschiedliches Bemühen um eine erfolgreiche Erinnerung und/oder eine anspruchsvollere Enkodierung des Feedbacks insbesondere bei falschen Antworten hindeuten und somit höchstwahrscheinlich auch Auswirkungen auf die Behaltensleistungen ausüben. Denn wie die signifikanten Korrelationen zwischen Zeit und Behalten nahelegen, fällt die Behaltensleistung mit wachsender Übungszeit höher aus.



## Vergleich der experimentellen Vokabelübungszeiten

Um eine möglichst hohe ökologische Validität anzustreben, sollten die Probanden ihre Übungszeit pro Übungsdurchgang vollständig selbst bestimmen, was unweigerlich ungleiche Lernzeiten zwischen den Übungsmethoden erwarten ließ. Insbesondere eine echte Short-Answer-Testung erfordert durch die schriftliche Eingabe der Antwort deutlich mehr Zeit als das reine Studium eines präsentierten Vokabelpaares (siehe analog auch Jacobs 2006). Es wurde daher angenommen, dass man in der Zeit eines konventionellen Testdurchgangs mit Feedback mindestens 2 mal alle Vokabeln durch Studieren bearbeiten kann und deshalb festgelegt, in der ersten Übungsphase 2 Durchgänge für das Studieren anzusetzen. Die Ergebnisse in Tabelle 5 bestätigen zunächst einmal die Annahme. Diese Regelung hätte man natürlich auch für die zweite Übungsphase übernehmen können, wenn lediglich SA gegen SO zum Vergleich angestanden hätte. Aber CSA liegt von der Bearbeitungszeit zwischen SA und SO. Eine vergleichbare Lernzeit beider Testmethoden mit SO lässt sich nicht erreichen. Aus pragmatischen Gründen wurde das Studieren daher in der zweiten Übungsphase auf einen Durchgang begrenzt in der Hoffnung, so eine annähernd vergleichbare Gesamtlernzeit gegenüber CSA zu erreichen. Wie aus Tabelle 5 hervorgeht, ging die Rechnung glücklicherweise auf.

Zunächst wurden die Zeiten aller Probanden mittels Tukey's Boxplot auf Ausreißer untersucht und fünf Ausreißer aus der Übungsphase 1, null aus der Übungsphase 2 und drei aus der Gesamtübungszeit herausgenommen. Tabelle 5 zeigt die Lernzeiten aller experimentellen Gruppen für alle Übungsphasen bzw. Übungszeitpunkte.

**Tabelle 5: Übungszeiten in Sekunden für alle experimentellen Bedingungen**

	Übung- phase 1 N = 30-33		Übungs- phase 2 N = 30-33		Übungszeit insgesamt N = 31-32	
	M	s	M	s	M	s
SO Studieren	140	71	59	56	<b>203</b>	117
CSA Testen mit KCR	129	75	93	61	<b>217</b>	124
SA Testen mit KCR	181	65	146	54	<b>324</b>	109

**Anmerkung:**

Übungsphase 1: CSA und SA umfassten jeweils einen Übungsdurchgang, SO zwei Übungsdurchgänge.

Übungsphase 2: Alle 3 Methoden umfassten jeweils einen Übungsdurchgang.

Die Korrelation der Zeiten zwischen den Übungsphasen 1 und 2 beträgt  $r=.63$  ( $N=93$ )

CSA sollte gegenüber SA schneller vollzogen werden, da die schriftliche Eingabe durch die gedankliche Antwort ersetzt wurde. Dieser ökonomische Vorteil bestätigt sich hier, da die Übungszeiten von CSA gegenüber SA sowohl im ersten ( $d=.75$ ) wie im zweiten Übungsdurchgang ( $d=.62$ ) signifikant geringer ausfielen ( $p$  einseitig jeweils  $<0.05$ ). Wie aus der zweiten Übungsphase hervorgeht, erforderte das Studieren der Vokabelpaare die geringste Bearbeitungszeit je Übungsdurchgang (signifikanter Zeitvorteil von Studieren gegenüber CSA und SA).

Für die Interpretation der Ergebnisse ist die Gesamtübungszeit von besonderer Relevanz, da Lernvorteile eigentlich nur bei vergleichbaren Lernzeiten klar interpretierbar sind. Hier ergab eine Testung aller möglichen Vergleiche folgendes Gesamtergebnis:  $(SO = CSA) < SA$  ( $< =$  signifikant geringer nach Bonferroni  $p<.001$ ). Wie die Zeiten zur Gesamtübungszeit offenlegen, fallen die Lernzeiten von Studieren und CSA sehr hoch vergleichbar aus. Das Studieren und die Testvariante CSA ermöglichen somit einen optimalen Vergleich ihrer Lernwirksam-

keit, während ein potenzieller Vorteil des echten Testens mit Feedback gegenüber dem Studieren teilweise zumindest auch der höheren Lernzeit angelastet werden könnte.

## Vergleich der Behaltenstests

Mindestens 3 bis höchstens 7 Tage nach der Übung bearbeiteten Probanden über Internet den ursprünglich eingeübten Behaltenstest sowie den Transferbehaltenstest. Da nicht alle Studierende zu diesen Tests angetreten waren, mussten einige Ausfälle (CSA = 4, SA = 7, SO = 4) hingenommen werden. In der nachfolgenden Seminarsitzung, ca. 1 bis 3 Tage nach den Online-Behaltenstests, wurden die Studierenden überraschend aufgefordert, alle Vokabelpaare aufzuschreiben, an die sie sich noch erinnern konnten. Weil im Seminar einige Studierende nicht anwesend waren, reduziert sich die Probandenanzahl für die entsprechenden Vergleiche weiter. In die folgende Analyse gehen nur solche Probanden ein, die den Versuch vollständig bearbeitet hatten, d.h. sowohl die Übung als auch die beiden Behaltenstests absolviert hatten. Nur bei diesen Probanden ging auch der free recall test in die Auswertung ein. Eine Analyse aller Behaltenswerte mittels Boxplot ergab keinerlei Ausreißer, obgleich auch Testwerte von 0 korrekten Vokabeln vorkamen.

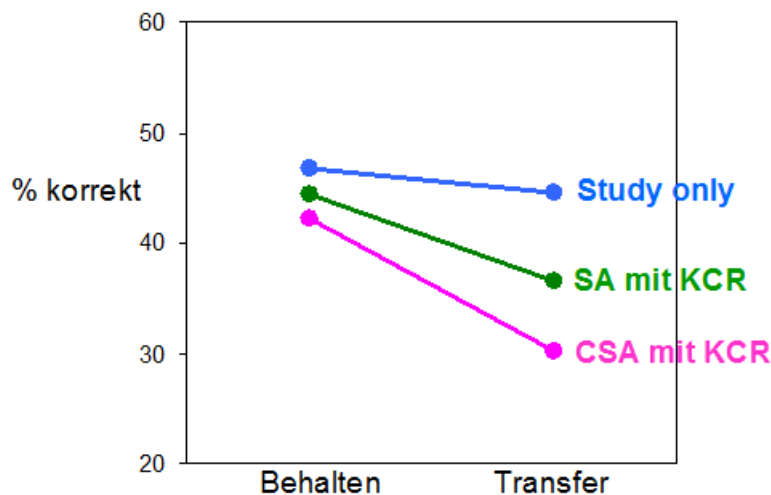
**Tabelle 6: Reliabilitäten und Interkorrelationen der Behaltenstests**

Behalten Transfer			
Cronbachs $\alpha$ :	.87	.88	N=82
Korrelationen			
Behalten Transfer			
Behalten	-	.84	N=82
free recall	.73	.71	N=68
Testzeiten	.49	.48	N=82

Tabelle 6 belegt die Zuverlässigkeit und relativ hohe Ähnlichkeit der erhobenen Behaltenswerte. Wer sich viele Vokabeln in der eingeübten Weise - "Wie heißt die deutsche Vokabel bei Vorlage der Fremdsprachenvokabel?" - merken kann, dem gelingt es auch sehr gut, die Fremdsprachenvokabel bei Vorlage der deutschen Vokabel zu erinnern sowie auch viele Vokabelpaare frei aus dem Gedächtnis zu reproduzieren. Beide Online-Behaltenstests zeigen, wie schon bei der SA-Übung, signifikante Zusammenhänge mit den Testbearbeitungszeiten in dem Sinne, dass längere Testbearbeitungszeiten beim Behaltens- und Transfertest mit entsprechend höheren Behaltenswerten einher gehen. Beim schriftlichen free recall test im Seminar konnten keine Bearbeitungszeiten erhoben werden. Tabelle 7 und Abbildung 2 stellen die Ergebnisse der Online-Behaltenstests dar.

**Tabelle 7: Prozentsatz korrekter Lösungen in den Behaltenstests**

	Behaltenstest			Transfertest	
	N	M	s	M	s
Studieren SO	28	<b>46.7</b>	28.6	<b>44.5</b>	28.6
CSA mit KCR	27	<b>42.2</b>	26.4	<b>30.2</b>	21.9
SA mit KCR	27	<b>44.4</b>	23.5	<b>36.5</b>	21.9

**Abbildung 2: Interaktion zwischen Übungsmethoden und Behaltensvariante**

Eine zweifaktorielle Varianzanalyse mit den Behaltensvarianten (Behaltenstest, Transfertest) als within subject factor und den experimentellen Übungsbedingungen (SO, CSA, SA) als between subject factor ergab einen deutlichen Effekt der Behaltensvarianten ( $F(1,79)=23.3$ ,  $p<.001$ ), keinen Haupteffekt der Übungsmethoden ( $F(2,79)=0.99$ ,  $p=0.38$ ), aber eine knapp signifikante Wechselwirkung zwischen Übungsmethoden und Behaltensvarianten ( $F(2,79)=3.36$ ,  $p=0.04$ ). Aus Tabelle 7 und Abbildung 2 geht klar hervor, dass die Übungsmethoden keine Unterschiede im eingeübten Behaltenstest erzielten. Die Interaktion deutet in die Richtung, dass die Übungsvarianten, welche eine Testkomponente enthalten, vornehmlich CSA, einen geringeren Transfer als Studieren bewirkten. Der einfache Mittelwertsvergleich des Transfertestes zwischen Studieren und CSA via t-Test für unabhängige Stichproben ergab einen signifikanten t-Wert ( $t(53)=2.05$ ,  $p=0.045$ ), der einer Effektstärke von  $d=.56$  entspricht. Der im Hauptfaktor Testbehaltensvarianten nachgewiesene Leistungsabfall vom Behaltens- zum Transfertest geht fast vollständig auf die Testübungsvarianten zurück. Denn während CSA und SA signifikant geringere Transfer- als Behaltenswerte aufweisen ( $t(26)=3.86$ ,  $p<.001$ ;  $t(26)=3.29$ ,  $p=0.003$ ), kann dieser Abfall bei SO nicht festgestellt werden ( $t(27)=0.95$ ,  $p=0.348$ ).

**Tabelle 8: Prozentsatz korrekter Lösungen im free recall test**

	N	M	s
Studieren SO	22	45.5	28.0
Testen CSA	25	37.0	20.1
Testen SA	21	37.3	20.8

Die Ergebnisse des free recall tests in Tabelle 8 deuten auf einen numerischen Vorteil des Studierens hin, der aber statistisch nicht gesichert werden kann. Weder der Haupteffekt der einfaktoriellen VA ( $F(2,65)=0.95$ ,  $p=0.39$ ), noch ein Kontrast: Studieren vs. beide Testübungsvarianten ( $t(65)=1.37$ ,  $p=0.17$ ) noch irgendeiner der möglichen Mittelwertsvergleiche ergab ein signifikantes Ergebnis. Auch hinsichtlich der Bearbeitungszeiten der Behaltenstests ließen sich keine signifikanten Unterschiede feststellen.

Alle Befunde widerlegen die Hypothese einer höheren Lerneffektivität der Testübungsmethoden gegenüber dem erneuten Studieren. Da Studieren und CSA hoch vergleichbare Übungszeiten aufwiesen, reicht die Lerneffizienz von CSA höchstens an die des Studierens heran und schneidet beim Transfertest sogar schlechter ab, ein deutlicher Widerspruch zu den Ergebnissen von Carpenter et al. (2006). SA-Testen umfasste zwar einen Übungsdurchgang weniger

als erneutes Studieren, erforderte aber dennoch mehr Gesamtübungszeit, so dass die Lerneffizienz des erneuten Studierens sogar etwas besser einzuschätzen ist als die des überprüfbaren Testens mit Feedback. Auch dieses Ergebnis widerspricht den bisherigen Befunden.

### Subjektive Leistungseinschätzungen der Studierenden im Vergleich zum objektiven Ergebnis

Unmittelbar vor dem Online-Behaltenstest sollten die Probanden einschätzen, wie viel Prozent der Vokabeln Sie vermutlich korrekt lösen werden (=Judgment of learning [Jol]), unmittelbar nach der Testung des Behaltenstests, wie viele Vokabeln sie vermutlich korrekt gelöst hatten (Judgement of performance [Jop]). Danach erhielten sie Rückmeldung über ihren tatsächlichen Prozentsatz der korrekten Lösungen und anschließend folgte der Transfertest. Jol und Jop beziehen sich folglich nur auf den Behaltenstest. Des Öfteren wird in der Test- und Feedbackforschung behauptet, Studieren begünstige Kompetenzillusionen, die beim Testen weniger zu erwarten seien. Allerdings setzen die Fragen meist unmittelbar nach der Übungsphase an und beziehen sich dann auf eine Schätzung zu einem späteren Zeitpunkt. Hier wird die Schätzung ja erst zum späten Zeitpunkt initiiert. In der nachfolgenden Analyse werden sowohl die Mittelwerte als auch die Korrelationen der subjektiven und objektiven Daten miteinander verglichen.

**Tabelle 9: Korrelationen zwischen subjektiver Leistungseinschätzung und objektivem Behaltenstest für alle Probanden (N=82)**

	Jol	Jop
Jol	-	.75
Behaltenstest	.62	.88

Die positiven Korrelationen zwischen der subjektiven Behaltenseinschätzung und dem objektiven Behalten belegen, dass sich die Probanden sowohl vor wie verständlicherweise noch deutlicher nach der Testung relativ zutreffend einschätzten. Tabelle 9 umfasst alle Probanden, weil bei allen experimentellen Bedingungen ähnliche Korrelationen festzustellen waren. Dass eine Testung in der Übung (SA oder CSA) bessere subjektive Leistungsprognosen ermöglicht hätte als erneutes Studieren, kann definitiv ausgeschlossen werden, weil die Korrelationen unter erneutem Studieren numerisch höher ausfallen. Korrelationen erfassen relative Übereinstimmungen aber keine Niveauunterschiede und geben daher keine Auskunft darüber, wie stark die einzelnen Schätzungen vom tatsächlichen Behalten abweichen.

**Tabelle 10: objektiver und subjektiver Prozentsatz korrekter Lösungen im Behaltenstest**

		Behaltenstest			Jol		Jop	
		N	M	s	M	s	M	s
Studieren	SO	28	<b>46.7</b>	28.6	<b>39.7</b>	27.0	<b>36.3</b>	31.8
Testen	CSA	27	<b>42.2</b>	26.4	<b>36.4</b>	19.7	<b>29.6</b>	25.3
Testen	SA	27	<b>44.4</b>	23.5	<b>28.2</b>	22.3	<b>25.7</b>	24.5

Zunächst fällt in Tabelle 10 auf, dass die Probanden im Durchschnitt ihre tatsächliche Leistung eher unterschätzten, was auf einen "underconfidence with practice effect" hinaus läuft (Koriat et al. 2002). Eine zweifaktorielle Varianzanalyse mit den Faktoren Leistung (objektiv, subjektiv [nur jol]) als within subject- und den Übungsmethoden als between subject factor erbrachte einen eindeutigen Effekt des Faktors Leistung zugunsten der objektiven Behaltensleistung ( $F(1,79)=16$ ,  $p<.001$ ), aber keine signifikante Interaktion mit den experimentellen Bedingungen ( $F(2,79)=1.94$ ,  $p=0.15$ ). Die numerisch größere Unterschätzung beim SA-Testen reicht demnach nicht aus, um sich klar gegenüber dem Studieren durchzusetzen. Auch

wenn Mittelwerte immerhin Auskunft über durchschnittliche Über- oder Unterschätzungen ermöglichen, so liefern sie dennoch kein hinreichendes Maß für die Genauigkeit einer Schätzung. Meiner Meinung nach wird die Genauigkeit am besten durch den Betrag der Abweichung von subjektiver Schätzung und objektiver Leistung erfasst. Eine statistische Analyse hinsichtlich des Betrags der Abweichung vom korrekten Ergebnis erbrachte aber ebenfalls keinerlei signifikante Unterschiede zwischen den experimentellen Gruppen. Alle Ergebnisse deuten darauf hin, dass die Übungsmethoden keine unterschiedlichen subjektiven Schätzungen bewirkten, die im Hinblick auf die Selbstregulation des Übens metakognitive Vorteile im Hinblick auf die Genauigkeitswahrnehmung des eigenen Wissens nahe legen würden.

## Zusammenfassung und Diskussion

Die bisherigen Studien des Verfassers zum Vergleich der Behaltenswirksamkeit zwischen Testen und Studieren erbrachten keinen überzeugenden Vorteil des Testens mit Feedback gegenüber dem wiederholten Studieren. Während Testen bei einfachstem Lehrzielniveau unter Labor ähnlichen Bedingungen meistens Vorteile erkennen ließ, blieb der strenge Testeffekt bei anspruchsvolleren Lehrzielen im Universitätssetting häufig aus (z.B. Jacobs, 2011). Nicht zuletzt deshalb wurde der Versuch unternommen, die in der experimentellen Forschung mehrfach bestätigte Überlegenheit des Testens an trivialem Faktenwissen nun deutlich bestätigen zu können, aber dabei auch einige realistische Lernbedingungen der pädagogischen Praxis zugrunde zu legen. Der Versuch führte zu unerwarteten Ergebnissen. Alle Übungsmethoden erzielten vergleichbare Behaltenswerte, womit höhere Transferwerte bei den Testvarianten schon gar nicht mehr erwartet werden konnten. Diese Ergebnisse widersprechen den Befunden von Carpenter et al. (2006) sowie Wei-chieh (2010), die von höheren Behaltens- und Transferwerten berichten. Hier fallen die Transferwerte von CSA sogar niedriger aus als die unter erneutem Studieren. Möglicherweise setzte sich die numerische Überlegenheit im Behaltentest bei SO erst im Transfer durch, zumal der dem Transfertest vorausgehende Behaltentest für SO die erste Testung darstellte und eventuell eine höhere Lerneffizienz bewirkte als die 3. Testung der Testgruppen. Für die Praxis empfiehlt es sich natürlich, bereits in den Übungen die Richtung der Vokabelabfrage zu variieren, um einer einseitigen Lernan-eignung vorzubeugen, was der Verfasser dieser Arbeit schon zu MS-Dos-Zeiten in einem [Vokabelprogramm](#) vorsah.

Im Labor können manche Prozesse strenger kontrolliert werden als bei einer Online-Studie, aber Schüler und Studenten lernen normalerweise zu Hause und nicht im Labor. Die höhere ökologische Validität des natürlichen Lernens spricht selbstverständlich nicht grundsätzlich gegen Laborforschung und wäre sowie wertlos, wenn es dem Vorgehen an interner Validität mangelte. Vorliegende Studie versteht sich als ein Bindeglied zwischen Labor- und Feldforschung, da sie einerseits einen größeren Freiheitsspielraum gewährte, aber auch etliche Kontrollen einführte, wie sie im alltäglichen Lernen gar nicht auftreten. So wurde das massive Lernen zugunsten von verteiltem Lernen quasi erzwungen. Das Erlernen von Swahilivokablen stand in keinerlei Zusammenhang mit dem Studium, war vermutlich weder von nennenswerter intrinsischer noch extrinsischer Motivation beflügelt worden und widerspricht diesbezüglich eindeutig einem Bemühen um externe Validität. Die Studie hat meiner Meinung nach aber deutlich gemacht, dass relativ konsistente Ergebnisse der Behaltensforschung im Labor nicht ohne Probleme auf die pädagogische Praxis übertragen werden können.

### **Erneutes Studieren eher besser als Testen mit Feedback bei vergleichbarer Übungszeit**

Positive Studien für die Überlegenheit des Testens, welche die Lernzeiten beider Bedingungen durch identische Itembearbeitungszeiten konstant gehalten hatten, kann hier der Vergleich SO gegen CSA entgegengestellt werden. Bei vergleichbaren Gesamtübungszeiten ließ sich im

Behaltenstest kein Unterschied nachweisen, aber beim Transfertest schnitt erneutes Studieren besser ab als CSA-Testen mit Feedback. Bei einigen, die Testung favorisierenden Studien verzichteten die Autoren auf die Erhebung oder Mitteilung von Übungszeiten. So fand etwa Karpicke (2007) in seinen Experimenten deutliche Vorteile fürs Testen, bestätigte in einer persönlichen Mitteilung aber meine Hypothese, die Testung hätte dann auch mehr Übungszeit benötigt. Manche mögen die Fairness des hier vorliegenden Methodenvergleiches bezweifeln, weil die Studiergruppe einen Übungsdurchgang mehr als die Testgruppen absolvierte. Aber zum einen erforderte dreimaliges Studieren insgesamt nicht mehr Zeit als zweimaliges Testen mit Feedback, zum anderen betrachten manche Forscher einmaliges Testen mit Feedback als zwei Übungsdurchgänge, weil ja sowohl ein Test stattfindet wie auch durch das Feedback ein Studieren ermöglicht wird. Allerdings wäre die Lernwirksamkeit höher einzuschätzen, wenn sich der Testung kein unmittelbares Feedback angeschlossen hätte, sondern auf einen reinen Testdurchgang der Studierdurchgang quasi als verzögertes Feedback gefolgt wäre (Butler et al. 2007, Jacobs 2008b).

Die Erwartung eines unmittelbaren Feedbacks birgt die Gefahr in sich, den anstrengenden Prozess des Erinnerns eher abubrechen und das Feedback einzufordern, was die Behaltenswirksamkeit des Testens schmälern könnte. Das gilt insbesondere für die Testmethode CSA. Die Erfolgsquoten während der SA-Übung und die annähernd vergleichbaren Zeiten für die Übungs- und die reinen Behaltenstests (ohne Feedback) unter der Bedingung SA schließen jedoch weitgehend aus, die Studierenden hätten diese Testübungsversion gehäuft als reine Studiermethode missbraucht.

### **Mangelndes Interesse als möglicher Einwand?**

Man wird nicht zu Unrecht vermuten, einige Studierende hätten den Versuch sehr lasch bearbeitet und seien gar nicht ernsthaft willens gewesen, Swahilivokabeln zu lernen. Diese These kann allerdings kaum als Argument gelten, da sie ja für alle Bedingungen galt, derartige Phänomene in jeder Untersuchung vorkommen und diese Konstellation im realen Schulsystem regelmäßig vorzufinden ist. Dessen ungeachtet wurden probenhalber alle Studierenden ausgeschlossen, die im Behaltens- oder Transfertest weniger als 10 Prozent der Aufgaben korrekt gelöst hatten. Dadurch änderten sich die entscheidenden Ergebnisse (etwa die von Abbildung 2) aber nicht grundsätzlich, sondern tendierten eher weiter in Richtung einer besseren Behaltensleistung durch Studieren im Vergleich zu beiden Testvarianten, die nach dem Ausschluss dieser Studierenden im Transfertest beide signifikant schlechter als Studieren abschnitten, sich untereinander aber nicht unterschieden.

### **Covert short Answer als alternative Übungsvariante zu klassischem Short Answer**

CSA war als ökonomische Alternative zu SA eingeführt worden und benötigte erwartungsgemäß signifikant weniger Übungszeit. Hinsichtlich aller Behaltenswerte ließen sich hingegen keinerlei statistische Unterschiede zwischen beiden Testmethoden nachweisen. Wenn Testen nur zu Übungszwecken zum Einsatz kommen sollte, wäre demnach die gedankliche Antwort als erwägenswerte Alternative zur schriftlichen Eingabe ins Kalkül zu ziehen. CSA bietet auch ohne aufwändige technische Hilfsmittel einen sehr flexiblen Übungseinsatz, z.B. auf der Basis von Karteikarten, die auf der einen Seite die Fremdsprachenvokabel und auf der anderen Seite die deutsche Vokabel zeigen, hat gegenüber SA allerdings den Nachteil, das Aufgabenergebnis mit einer etwas geringeren Objektivität zu bewerten und dem Studierenden keine verlässliche Auskunft seines aktuellen Wissensniveaus, etwa den Prozentsatz der korrekten Lösungen, vermitteln zu können.

### Abschließende Bemerkungen

Angesichts der Befunde macht es wenig Sinn, über diverse Theorien zu spekulieren, warum Testen eine höhere Behaltensleistung bewirken soll als erneutes Studieren. Die Bedingungen für das Testen waren insofern nicht optimal, als die Erfolgsquoten bei der ersten Testung mit 50% und die der zweiten Testung mit 60% vielleicht zu gering ausfielen. Denn die positive Wirkung reinen Testens kann sich nur bei erfolgreichem Retrieval zeigen. Die Präsentation des Feedbacks unterscheidet sich nicht sonderlich vom Studieren. Die Überlegenheit des reinen Testens gegenüber dem Studieren kann erst ab einer gewissen Testerfolgsquote einsetzen. Allerdings fanden manche Forscher langfristige Behaltensvorteile des Testens bei ähnlichen und selbst bei geringeren Erfolgsquoten als sie hier gefunden wurden.

Im Idealfall sollte der Abruf aus dem Gedächtnis anstrengend und dennoch erfolgreich sein, um das Behalten optimal zu stärken (Pyc & Rawson, (2009)). Deshalb wurde hier im Ansatz ein eher verteiltes Lernen angestrebt. Die geistig anspruchsvollen Aktivitäten zwischen den Lern- bzw. Übungsphasen sollten allzu leichtem Erinnern der Vokabeln entgegen wirken. Möglicherweise fördert verteiltes Lernen das erneute Studieren aber mindestens so stark wie das Testen.

Wie die durchwegs hohen Reliabilitäten der Vokabeltests und die marginalen Unterschiede der Übungsmethoden nahelegen, geht fast die gesamte Behaltensvarianz auf die Unterschiedlichkeit der Personen zurück, was im Übrigen die häufige Verwendung von Wiederholungsexperimenten in diesem Bereich nur allzu verständlich macht. Jacobs fand eine solche Konstellation in vielen seiner Experimente. Die Ergebnisse legen folgende Interpretation nahe: Die Präsentation eines Vokabelpaares zum Studieren und das Testen mit anschließendem Feedback gewähren vergleichbar gute Chancen zum Einprägen und Behalten. Welches Behaltensergebnis dabei herauskommt, hängt im Wesentlichen von der Fähigkeit und der Lernbereitschaft der Person ab, die jeweilige Übungsgrundlage angemessen zu nutzen. Gedankenlose Kenntnisnahme beim Studieren führt genau wenig zum Erfolg wie hastiges Testen und oberflächliche Betrachtung des Feedbacks, wie andererseits gründliches Überdenken und tiefe Elaborationen sowohl beim Testen wie beim Studieren förderlich sind. Die adäquate Nutzung beider Instruktionsvarianten setzt allerdings hinreichende Freiheit voraus, sich ohne rigorose Zeiteinschränkung oder sonstiger Gängelung den Aufgaben widmen zu können.

Testen mit Feedback verspricht einige motivationale Vorteile gegenüber reinem Studieren. In Übungssituationen ohne Leistungsbewertung präferieren Studierende Testen mit Feedback gegenüber wiederholtem Studieren und schätzen die pädagogische Qualität des Testens häufig höher ein (z.B. Jacobs 2010). Zudem bieten Tests die Option, den Studierenden zutreffende Rückmeldungen über ihre erzielte Leistung zu gewähren, wodurch diese nicht auf eigene Schätzungen ihres aktuellen Kenntnisstandes angewiesen sind. Tests ermöglichen darüber hinaus eine gezielte Aufgabensteuerung über geeignete Flashcardmethoden. Schließlich lassen sich Testergebnisse wesentlich einfacher und gezielter mit extrinsischen Konsequenzen belegen und so manche Studierende zu einem ernsthafteren Übungsverhalten bewegen.

### Literatur

- Butler, A. C., Karpicke, J. D. & Roediger, H. L. (2007). The Effect of Type and Timing of Feedback on Learning from Multiple-Choice Tests. *Journal of Experimental Psychology: Applied*, 13 (4), 273-281.
- Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13, 826-830.
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name-learning. *Applied Cognitive Psychology*, 19, 619-636.

- Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20, 632-642.
- Cranney, J., Ahn, M., McKinnon, R., Morris, S. & Watts, K. (2009): The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting, *European Journal of Cognitive Psychology*, 21(6)919-940
- Cull, W. L. (2000). Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall. *Appl. Cognit. Psychol.* 14: 215-235.
- Hamaker, Ch. (1986). The Effects of Adjunct Questions on Prose Learning. *Review of Educational Research*, 56 (2) 212-242.
- 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. (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). Die Behaltenswirksamkeit wiederholten Einprägens im Vergleich zu Computer- und selbst gesteuertem Testen mit Feedback.  
URN: urn:nbn:de:bsz:291-psydok-26913  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2010/2691/>
- Jacobs, B. (2008). Führt selbst gesteuertes Testen mit Feedback zu höheren Behaltensleistungen als das Einprägen mit Hilfe einer Landkarte?  
URN: urn:nbn:de:bsz:291-psydok-27625  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2011/2762/>
- Jacobs, B. (2008b). Einige Überlegungen zu unmittelbarem und verzögertem Feedback.  
URL: [http://www.phil.unisb.de/~jakobs/wwwartikel/feedback/unmittelbar\\_vs\\_delayed.html](http://www.phil.unisb.de/~jakobs/wwwartikel/feedback/unmittelbar_vs_delayed.html)  
[1.7.2011]
- Jacobs, B. (2009). Die Wirkung der Übungsverteilung beim Studieren und Testen auf das Behalten der Bundesstaaten der USA.  
URN: urn:nbn:de:bsz:291-psydok-23544  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2009/2354/>
- Jacobs, B. (2010). Testfragen selbst beantworten oder Musterlösungen studieren?  
URN: urn:nbn:de:bsz:291-psydok-26934  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2010/2693/>
- Jacobs, B. (2011). Musterlösungen durcharbeiten als Alternative zu Testen mit Feedback - Eine Replikationsstudie.  
URN: urn:nbn:de:bsz:291-psydok-27127  
URL: <http://psydok.sulb.uni-saarland.de/volltexte/2011/2712/>
- Karpicke, J. D. (2007). STUDENTS' USE OF SELF-TESTING AS A STRATEGY TO ENHANCE LEARNING. Dissertation. Graduate School of Arts and Sciences of Washington University. Saint Louis, Missouri.
- Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology Learning, Memory, and Cognition*, 33 (4), 704-719.
- Karpicke, J. D. & Roediger H. L. (2008) The critical importance of retrieval for learning. *Science* 319, 966-968.
- Koriat, A., Sheffer, L., Ma'ayam, H. (2002). Comparing objective and subjective learning curves: Judgements of learning exhibit increased underconfidence with practice. *Journal of experimental psychology*, 131, 147-162.
- Pyc, M.A., Rawson, K.A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language* 60, 437-447.
- Pyc, M.A., Rawson, K.A. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science*, 330, 335.
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103, 399-414.



- Nachtigall, C. & Wolf, A. (2001). Fragebogen zur Wahrscheinlichkeitstheorie (FWT)  
URL: [www.metheval.uni-jena.de/materialien/reports/report\\_2001\\_03.pdf](http://www.metheval.uni-jena.de/materialien/reports/report_2001_03.pdf) [20.12.2011]
- Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249-255.
- Schmidmaier, R., Ebersbach, R., Schiller, M. Hege, I., Holzer, M. & Fischer, M. R. (2011). Using electronic flashcards to promote learning in medical students: retesting versus restudying. *Medical Education*, 1101-1110
- Toppino T. C, Cohen M. S. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology*, 56 (4), 252-257.
- Vilhelmsson, S. (2011). Learning material with different difficulty: When is testing more beneficial than study? Bachelor Thesis in Cognitive Science. Umeå University Department of Psychology  
URL: <http://umu.diva-portal.org/smash/get/diva2:453211/FULLTEXT01> [20.12.2011]
- Wei-chieh Fang (2010). Testing Effects: Promote Transfer of Learning. Master Thesis. National Taiwan University of Science and Technology  
[http://140.118.33.1/ETD-db/ETD-search/view\\_etd?URN=etd-0720110-011011](http://140.118.33.1/ETD-db/ETD-search/view_etd?URN=etd-0720110-011011)  
<http://140.118.33.1/ETD-db/ETD-search/getfile?URN=etd-0720110-011011&filename=etd-0720110-011011.pdf>

## Anhang

### Vokabeln des Experimentes.

barua Brief  
bata Ente  
chakula Nahrung  
gari Auto  
hewa Luft  
jambo Thema  
jengo Gebäude  
kiwanda Fabrik  
matunda Obst  
maziwa Milch  
moto Feuer  
paka Katze  
pombe Alkohol  
ramani Landkarte  
sanduku Koffer  
sarafu Münze  
subira Geduld  
tatizo Problem  
wali Reis  
waridi Rose

## Ausschnitte aus Bildschirmkopien der experimentellen Varianten

Vokabel einprägen und dann nächste Vokabel anfordern!

Studieren  
Study Only  
SO

barua

Brief

nächste Vokabel anfordern

---

Deutsche Vokabel ins Gedächtnis rufen, leise aussprechen und dann zur Kontrolle auf 'Korrekte Antwort' klicken!

barua

Korrekte Antwort

Deutsche Vokabel ins Gedächtnis rufen, leise aussprechen und dann zur Kontrolle auf 'Korrekte Antwort' klicken!

barua

Brief

nächste Vokabel anfordern

Covert Short Answer (CSA)

---

Short Answer (SA)

Fremdsprache	Deine Antwort
barua	
Korrekte Antwort	Antwort bestätigen

---

Fremdsprache	Deine Antwort
barua	Ente <span style="color: red; font-size: 1.5em;">✗</span>
Korrekte Antwort	Brief

nächste Aufgabe anfordern