B. Jacobs & J. Sparfeldt, Bildungswissenschaften der Universität des Saarlandes
Version vom 6.10.2014

Musterlösungen und Testen mit Feedback als vergleichbar lernwirksame Übungsmethoden in der universitären Lehre

Abstract

Im Mittelpunkt der Studie steht ein empirischer Vergleich des Lernerfolgs zweier Übungsmethoden, welche dieselben Informationen beinhalten, sich aber in der Art des Zugangs zu den Informationen unterscheiden. Während das Testen zunächst eine aktive Antwort des Lernenden einfordert und erst danach die Rückmeldung folgt, sieht der Lernende bei einer Musterlösung direkt die korrekte Antwort und soll sich diese verständlich machen. Im Gegensatz zur Laborforschung, die häufig den Vorteil des Testens, den sogenannten strengen Testeffekt [testing effect], nachgewiesen hat, lag hier die vom Erstautor bereits mehrfach belegte Hypothese zugrunde, im realen Universitätsbetrieb bewirkten beide Übungsmethoden einen vergleichbaren Lernerfolg, insbesondere dann, wenn die Übungen Lehrziele eines vorausgegangenen Unterrichts stärken und hierbei besonderen Wert auf Verstehen und Anwendung legen. In einem klassischen Randomisierungsdesign mit Wiederholungsmessung ergaben sich erwartungsgemäß vergleichbare Lernerfolgswerte für beide Übungsmethoden bei teilweise kürzeren Bearbeitungszeiten für die Musterlösung, obgleich Studierende die pädagogische Qualität von Quiz mit Feedback eindeutig höher bewerteten als die der Musterlösungen. Bei den Quiz spielte es keine Rolle, ob das Feedback unmittelbar nach jeder Aufgabe oder erst am Ende der gesamten Quizbearbeitung gewährte wurde. Beide Quizvarianten erlaubten im Vergleich zu den Musterlösungen aber keine bessere Einschätzung des eigenen Wissens (Judgement of learning). Insgesamt deuten die Befunde darauf hin, Musterlösungen als wertvolle Übungsalternative zu Testen mit Feedback zu betrachten.

Schlagworte: Testen, Feedback, Musterlösung, Quiz, Feedbacktiming, Übung, „testing effect“, Hochschuldidaktik

Einleitung und Zielsetzung

Testen vs. Musterlösung

Testen dient nicht nur diagnostischen Zwecken zur Erfassung des Wissens, sondern stärkt das Behalten. Folgt dem Testen ein informatives Feedback, so vermag dieses neben der Stabilisierung korrekter Antworten hauptsächlich zur Korrektur von Fehlern genutzt werden und verbessert in der Regel die Lernwirksamkeit gegenüber reinem Testen. Testen mit Feedback im Anschluss an eine Instruktion kann somit als effektive Übung des zuvor vermittelten Lehrstoffs interpretiert werden. Während empirische Studien die Lernwirksamkeit des Testens mit und selbst ohne Feedback gegenüber einer No-Treatment-Kontrollgruppe hinreichend belegen, bleibt die Frage offen, welche sonstigen Übungsmethoden ebenso oder gar besser als Testen geeignet wären, den Lernerfolg zu fördern.

In diesem Zusammenhang wurde als Alternative zum Testen mit Feedback häufig das gezielte Studieren der Test relevanten Information herangezogen. Letztlich geht es hierbei um den Vergleich Testen vs. erneutes Studieren (Restudying, Review, Study only usw.), das in mehreren Varianten zur Anwendung kommt. Beim Vokabellernen werden die Vokabelpaare erneut zum Lernen vorgelegt. Aus einer Frage mit kurzer Antwort wird ein Aussagesatz, der die korrekte Antwort enthält (read statement nach LaPorte & Voss, 1975). Eine knappe Textpassage fasst die wichtigsten Aussagen zusammen, die ein Testkandidat eigenständig generieren muss. Statt selbst Fragen zu beantworten, wird etwa der Lehrtext ein zweites Mal zum Einprägen präsentiert. Allerdings umfasst das Studieren dann mehr Information als das Testen, weil in der Re-

gel nicht alles getestet wird. Im Idealfall sollte das Studieritem aber exakt dem korrekt beantworteten Testitem entsprechen und ein solches Studieritem bezeichnen wir im Folgenden als Musterlösung. Zur Musterlösung gehören somit die Fragestellung, die möglichen Distraktoren einer MC-Aufgaben, die korrekte Antwort und im Falle einer ausführlichen Rückmeldung auch der elaborierte Teil des Feedbacks.

Manche Testforscher erwarten beim Vergleich Testen vs. erneutem Studieren höhere Behaltensleistungen für das Testen mit der Begründung, Testen verlange einen anstrengenden Abruf aus dem Gedächtnis, bahne so mehr Assoziationen zur korrekten Lösung und erleichtere auf diese Weise den erneuten Zugriff auf das Wissen. Etliche Laborstudien haben die Überlegenheit des Testens mit und teilweise sogar ohne Feedback gegenüber der Präsentation einer Musterlösung nachgewiesen (z.B. LaPorte & Voss, 1975; Carrier & Pashler, 1992; Cull, 2000; Jacobs, 2006; Kang, McDermott & Roediger, 2007; Toppino & Cohen, 2009; Rohrer, Taylor & Sholar, 2010; Butler, 2010; für einen Gesamtüberblick siehe die Metaanalyse von Rowland, 2014). Auch Studien in realen schulischen Umwelten konnten den Testvorteil gegenüber einer Musterlösung belegen (z.B. Larsen, Butler & Roediger, 2009; McDaniel, Anderson, Derbish & Morrisette, 2007; Roediger, Argawal, McDaniel & McDermott, 2011 (Experiment 2), McDaniel, Wildman & Anderson (2012). Während beim unmittelbaren Lernerfolgstest die Musterlösung teilweise mehr Behalten als das Testen bewirkte, zeigte sich der Behaltensvorteil des Testens gegenüber einer Musterlösung in der Regel nur nach einem längerem Rententionsintervall (z.B. Roediger & Karpicke, 2006b; Toppino & Cohen, 2009). Der erwartete längerfristige Behaltensunterschied fiel aber gelegentlich auch recht mager aus, bezog sich meistens auf identische Aufgaben in Übung und Lernerfolgstest und zeigte sich vornehmlich bei Aufgaben des Typs constructed response (short answer oder free recall test).

Besonders schwierig erscheint es, die Behaltensvorteile des Testens mit Feedback bei Multiple Choice- Übungsaufgaben nachzuweisen (siehe jedoch Roediger et al., 2011, Experiment 2). In etlichen Fällen ist dies weder in Laborstudien (z.B. Butler & Roediger, 2007); Teilergebnisse aus den Studien von Kang et al., 2007), noch in realen pädagogischen Umwelten gelungen (z.B. Pilotti, Chodorow & Petrov, 2009; Howard, 2010). Laut einer Metaanalyse von Rowland (2014) lässt sich der strenge testing effect jedoch auch bei MC-Aufgaben belegen, fällt aber geringer aus als bei den constructed response Aufgaben. Eine theoretische Erklärung für die schwächere empirische Evidenz des Testeffekts bei MC-Aufgaben wird darin vermutet, das Bemühen um behaltenswirksame Prozesse, die Lösung aus dem Gedächtnis abzurufen, sei bei MC-Aufgaben weniger gefordert, da dort mehr Hinweise zur korrekten Antwort vorgegeben würden. Diese verlangten lediglich etwa bei Faktenwissen Wiedererkennung. Was für Faktenwissen und ganz einfache Verstehensleistungen zutreffen mag, gilt aber nicht für Aufgaben höheren Lehrzielniveaus, wie z.B. Anwendung. Schließlich erfordern MC-Aufgaben eines Intelligenztestes ja auch kein simples Wiedererkennen der korrekten Antwort. Das gebundene Aufgabenformat von MC- sowie weitere geschlossene Aufgaben ermöglichen aber eine objektive Auswertung und bieten sich insbesondere für eine, via Computer ökonomisch handhabbare, häufige Testung in der pädagogischen Praxis an.

Wie ein Vergleich "Testen mit Feedback vs. Musterlösung" ausfallen wird, hängt vermutlich auch entscheidend vom Lehrzielniveau und der Schwierigkeit der Aufgaben ab. In der Forschung zum „testing effect“ überwiegen einfache Lehrziele wie Faktenwissen oder Textverstehen. Die Probanden erzielten bei der Testung meistens recht

hohe Erfolgsquoten, was nach der Metaanalyse von Rowland (2014) die Wirkung des testing effects erhöht. Die Worked Examples Forschung, zu Deutsch Forschung von Lösungsbeispielen, untersuchte demgegenüber vorwiegend eher Problemlöseaufgaben (z.B. Atkinson et al., 2000). Ihre Ergebnisse weisen klar in die Richtung, in einem initialen Lernstadium sei die Bearbeitung von Lösungsbeispielen (Worked Examples) deutlich lerneffizienter als eigenständige Lösungsversuche (=Testung), selbst dann, wenn nach der Aufgabenbearbeitung elaboriertes Feedback folgt (z.B. Paas & Van Merrienboer, 1994). Auch fielen die Erfolgswahrscheinlichkeiten bei eigenen Problemlösungen in der Regel ziemlich gering aus.

Das hier im Rahmen einer Diagnostikvorlesung für Lehramtskandidaten angesiedelte Lehrzielniveau der Quiz umfasst unterschiedliche Anforderungen, darunter aber sehr selten reines Faktenwissen, sondern überwiegend Verstehen und Anwendung, jedoch keine Problemlösung. Die bisherigen Studien des Erstautors bezogen sich ebenfalls auf das Themengebiet Pädagogische Diagnostik, nutzten die Möglichkeiten des Internets für elektronisch vermittelte Übungen und führten fast ausnahmslos zu vergleichbaren Lernerfolgen beider Übungsvarianten bei gleichzeitig geringeren Lernzeiten für die Musterlösungen. Da hier ähnliche Aufgabentypen, Anforderungen und Themen zur Anwendung kommen, erwarten wir auch für vorliegende Untersuchung

a. vergleichbare Lernerfolgsergebnisse ca. eine Woche nach den Übungen sowie
b. kürzere Bearbeitungszeiten für die Musterlösungen.

Darüber hinaus erbrachten die bisherigen Studien von Jacobs (2010, 2011) zur Wertschätzung beider Übungsmethoden übereinstimmend deutliche Präferenzen für das Testen mit Rückmeldung gegenüber der Präsentation einer Musterlösung, aber keine bessere Einschätzung des eigenen Wissens. Ähnliche Ergebnisse werden deshalb auch in dieser Studie erwartet.

Timing des Feedbacks

Nebenbei sollte die Untersuchung klären, ob der Zeitpunkt der Rückmeldungen bei den Quiz einen Einfluss auf den Lernerfolg oder die Wertschätzung ausübt. Hierbei kamen zwei in der Praxis häufig anzutreffende Feedbacktimingmethoden zum Einsatz. Bei der einen Variante folgte die Rückmeldung unmittelbar nach jeder Aufgabe (FAI= **F**eedback **a**fter **I**tem), im anderen Fall unmittelbar nach Bearbeitung aller Aufgaben bzw. nach "Abgabe des Quiz" (EOTF=**E**nd **o**f **T**est **F**eedback). Die erste Methode setzt in der Regel den Einsatz des Computers voraus, während die zweite Variante auch ohne Computer leicht zu realisieren ist. Bisherige empirische Ergebnisse zum Vergleich beider Feedbackvarianten liefern ein uneinheitliches Bild. So fanden etwa Buzhardt & Semp (2002), Henshaw (2011) und van der Kleij et al. (2012) hoch vergleichbare Posttestergebnisse für FAI und EOTF, während Dihoff et al. (2003), Dihoff et al. (2004) und Brosvic & Epstein (2007) konsistent bessere Behaltenswerte für unmittelbares Feedback nach jeder Aufgabe feststellten. Bei den letztgenannten Autoren kam als unmittelbares Feedback eine spezielle Form des Answer until correct Feedbacks zum Einsatz, bei dem der Studierende im Fehlerfall die jeweils verbliebenen Alternativen der MC-Aufgabe anwählen sollte, bis er zur korrekten Lösung gelangte (IAT-Technik). Beim End of Test Feedback, also nach Abgabe des Quizzes, erhielt der Studierende seine Bearbeitung sowie eine Liste der korrekten Alternativen. Es wurde ihm dann in einer Reviewsitzung Zeit eingeräumt, die Rückmeldungen einzusehen, wobei

aber unklar blieb, wie ernsthaft Studierende von dieser Möglichkeit Gebrauch machten.

Während in der Laborforschung verzögertes Feedback meistens bessere Behaltensergebnisse erbrachte als unmittelbares Feedback (z.B. Butler et al., 2007), sprechen die Befunde in der Schulpraxis eher für das Gegenteil. Dort erwies sich unmittelbares Feedback als wirksamer (Kulik & Kulik, 1988), vermutlich, weil Schüler verzögertem Feedback weniger Aufmerksamkeit schenken. In der Forschungsliteratur ist der Begriff "verzögertes Feedback" nicht einheitlich durch einen bestimmten Mindestzeitabstand zur Testung definiert worden, sondern sehr unterschiedlich meist relativ zum Zeitpunkt unmittelbaren Feedbacks. In entsprechenden Untersuchungen kann verzögertes Feedback von einigen Sekunden bis Wochen variieren, da es im Prinzip nur später als unmittelbares Feedback folgen musste. EOTF nimmt unserer Meinung nach eine Art Mittelstellung zwischen unmittelbarem und verzögertem Feedback ein. Von echtem verzögertem Feedback würden wir eher sprechen, wenn Testung und Rückmeldung zu zwei unterschiedlichen Terminen stattfänden, die mindestens 24 Stunden auseinander liegen. Die möglichen Lernerfolgsunterschiede zwischen FAI und EOTF hängen vermutlich davon ab, wie ernsthaft die Testung in beiden Varianten vorgenommen und wie aufmerksam das Feedback jeweils gelesen werden und in welcher Funktion das Feedback genutzt wird, z.B. im Sinne einer Stoffwiederholung oder als Verständnishilfe bei Fehlern. Da die bisher bekannten empirischen Befunde keine eindeutige Aussage zulassen und manche theorierelevante Informationen hier nicht erfassbar sind, liegt hinsichtlich der Behaltensüberlegenheit einer der beiden Feedbackvarianten keine gerichtete Hypothese vor, wiewohl insgesamt mit eher geringen Unterschieden gerechnet wird, da die Effekte auch von dem durch Feedback überhaupt erreichbaren Lernpotenzial abhängen.

Empirische Untersuchung

Probanden, Untersuchungsablauf und Versuchsplan.

Als Probanden dienten Studierende des Lehramts der Universität des Saarlandes, die an der Vorlesung "Pädagogische Diagnostik und Intervention WS12/13" teilnahmen. Die Vorlesung war für alle Lehramtsstudierenden verpflichtend. Ca. 500 Studierende (ca. 1/3 männlich, 2/3 weiblich, Alter ca. 22 Jahre) hatten sich für die Vorlesung angemeldet.

In der ersten Vorlesungssitzung erging an die anwesenden Studierenden das Bonusangebot von wenigen Prozentpunkten für die Abschlussklausur, wenn Sie sich bereit erklärten, im Verlauf des Semesters an einigen wissenschaftlichen Erhebungen teilzunehmen, die zum Teil online ablaufen oder die Anwesenheit in der Vorlesung erfordern würden. Ca. 350 Studierende erklärten anfangs ihre Zusage, aber aus unterschiedlichen Gründen wie Krankheit, Vergesslichkeit, mangelnde Onlineerreichbarkeit usw. ergaben sich im Verlauf der Untersuchung etliche Ausfälle. Die Studierenden wurden im Vorfeld des Experimentes nicht darüber informiert, was sie erwartet. Ihnen wurde lediglich mitgeteilt, in der Zeit vom 14. bis 18.11.2012 sollten sie einen Arbeitsauftrag online erledigen. In diesem Zeitintervall fanden die experimentellen Übungen via Internet statt. Für den 26.11.2012 wurde eine weitere Erhebung in der Vorlesung angekündigt und dort dann überraschend der unbenotete Lernerfolgstest in der Papier und Bleistiftversion zur Bearbeitung vorgelegt.

Inhaltlich beziehen sich die Übungen auf den Lehrstoff der beiden vorausgegangenen Vorlesungen "Skalen und Bezugnormen (verkürzt Skalen)" und "Testgütekriterien (Gütekriterien)". Auf die Folien der entsprechenden Vorlesungen hatten die Studierenden bereits vor dem jeweiligen Vorlesungstermin Zugriff via Internet. Etliche Studierende brachten die ausgedruckten Folien mit in die Vorlesung, um sich entsprechende Notizen zu machen. Standardmäßig begann eine Vorlesungssitzung mit einer ca. 20 Minuten dauernden Wiederholung der wichtigsten Punkte der vorausgegangenen Vorlesungssitzung, so dass die Lernaneignungsphase für alle Studierenden eine umfangreiche und konzentrierte Darbietung des relevanten Lehrstoffs umfasste, der durch die Übungen dann gefestigt werden sollte. Das gesamte Vorgehen lässt sich wie in Tabelle 1 zusammenfassen.

Tabelle 1: Ablauf der Untersuchung

5.11.2012	12.11.2012	14.11 - 18.11.212		26.11.2012
Vorlesung Skalen	Vorlesung Gütekriterien	Online-Übung zu Hause Übung 1 Skalen	Übung 2 Gütekriterien	Lernerfolgstest zu Übung 1 und 2

Der experimentelle Teil bezieht sich auf die Online-Übung zuhause. Zunächst wurden alle verfügbaren Studierenden nach Zufall auf fünf Gruppen, vier experimentelle- und eine Kontrollgruppe, aufgeteilt. Entsprechend der ausgewählten Bedingung erhielt jeder Studierende eine Email mit der Internetadresse, Username und Passwort für "seinen Arbeitsauftrag". In dieser Email sowie im Vorspann der jeweiligen Bedingung standen mehrere Empfehlungen für ein ordentliches Arbeiten (z.B. „nur ein Fenster im Browser benutzen, Übung in einem Zug durcharbeiten, sonstige Programme oder Geräte wie Facebook, Google, Email, Handy abschalten usw.)

Alle experimentellen Gruppen bearbeiteten die zwei Übungen direkt hintereinander und zwar in Form einer der nachfolgenden Treatmentvarianten:

- Quiz mit unmittelbarer Rückmeldung nach jeder bearbeiteten Aufgabe (**Q_FAI**=**Q**uiz mit **F**eedback **a**fter **i**tem)
- Quiz mit Rückmeldung nach Bearbeitung aller Aufgaben (**Q_EOTF** = **Q**uiz mit **E**nd **o**f **T**est **F**eedback)
- Musterlösung = die korrekt beantworteten Quizaufgaben (**MU** = **Mu**sterlösung)

Wie aus Tabelle 2 hervorgeht, galt für alle experimentellen Gruppen, dass eine Übung als Musterlösung und die andere Übung als Quiz bearbeitet werden musste, und zwar das Quiz entweder mit unmittelbarer Rückmeldung oder erst nach Bearbeitung aller Aufgaben.

Tabelle 2: Versuchsplan

	G	Übung1	Übung2
R	1	FAI_1	MU_2
R	2	MU_1	FAI_2
R	3	EOTF_1	MU_2
R	4	MU_1	EOTF_2
R	5	KG_1	KG_2

```
Zeichenerklärung
Q_FAI_1  = Quiz mit Feedback after Item in Übung 1
Q_FAI_2  = Quiz mit Feedback after Item in Übung 2
Q_EOTF_1 = Quiz mit End of Test Feedback in Übung 1
Q_EOTF_2 = Quiz mit End of Test Feedback in Übung 2
MU_1     = Musterlösung in Übung 1
MU_2     = Musterlösung in Übung 2
KG_1     = Kontrollgruppe (Konzentrationstests) für Übung 1
KG_2     = Kontrollgruppe (Konzentrationstests) für Übung 2
R= Zufällige Zuweisung des Studierenden zu dieser Bedingung
G= Gruppe 1 bis 5
```

Für die erste Übung liegt ein klassischer experimenteller Versuchsplan zu Grunde, der die Prüfung der reinen Treatmenteffekte auf der Basis unabhängiger Vergleiche erlaubt. Er lässt sich weitgehend auch auf die Übung 2 übertragen, allerdings mit der Einschränkung, dass zu diesem Zeitpunkt zuvor bereits unterschiedliche Erfahrungen gemacht wurden. Da jede experimentelle Gruppe sowohl ein Quiz mit Feedback als auch eine Musterlösung bearbeitete, ermöglicht der Versuchsplan auch eine Abschätzung bestimmter Effekte auf der Basis eines Wiederholungsdesigns, wobei aber die Übungsthemen variierten. Die Reihenfolge der Übungsthemen war konstant vorgegeben, weil die Übung 1 (Skalen) stets vor der Übung 2 (Gütekriterien) stattfand. Übungszeitpunkt und Übungsthema sind somit konfundiert. Durch einen Längsschnittvergleich von Gruppe 1 und 2 bzw. 3 und 4 wurde das Übungsthema balanciert, nicht jedoch die Übungsreihenfolge.

Die Kontrollgruppe absolvierte vier kurze Konzentrationstests und dient hier für beide Übungen als Prüfstein, ob die speziellen experimentellen Übungen überhaupt einen spezifischen Lerneffekt hinterlassen würden. Da am Lernerfolgstest auch viele Studierende teilnahmen, die gar nicht zur Untersuchungsstichprobe gehörten, bot sich die Möglichkeit an, diese als weitere Kontrollgruppe zu verwenden. Diese No-Treatment-Kontrollgruppe - hier mit KG' bezeichnet - ist im Versuchsplan nicht aufgeführt, da ihr Zustandekommen nicht durch Randomisierung kontrolliert ist.

Effizienz der Randomisierung

Auf der Basis verfügbarer Daten der Probanden aus der Vorerhebung wurde die Vergleichbarkeit der fünf experimentellen Gruppen geprüft und bei Übungsantritt keine Unterschiede hinsichtlich möglicher relevanter Variablen gefunden. Varianzanalysen mit den fünf Treatmentgruppen als UV erbrachten keinerlei signifikante Haupteffekte hinsichtlich Abiturnotendurchschnitt, Alter, Geschlecht oder Angst vor der Diagnostikklausur. Keiner der möglichen Mittelwertsunterschiede zwischen allen Gruppen erreichte ein signifikantes Niveau von 5% nach LSD-multiplem Mittelwertvergleich, obwohl man nach Zufall schon mindestens einen hätte erwarten können. Ca. 80% aller Studierenden der Übungsgruppen nahmen eigenen Angaben zufolge an den relevanten Vorlesungen teil und ca. 70% hatten die Folien gelesen. Auch bzgl. des Vorlesungsbesuchs und des Lesens der relevanten Folien konnten keine Unterschiede zwischen Übungsgruppen festgestellt werden. Insgesamt muss die Effizienz der Randomisierung als hervorragend eingestuft werden.

Aufgaben und Charakteristika aller Übungen

Ziel der Übungen war es, das Verständnis für die in der Vorlesung behandelten Themen und nicht nur das Behalten des zuvor Dargebotenen oder Eingeübten zu stärken. Deshalb beinhalteten Übungs- und Lernerfolgsaufgaben nicht identische Items, sondern parallele Aufgaben, die teilweise zumindest einen geringen Transfer erfordern (siehe Beispiele unten).

Jede Übung war auf einer HTML-Seite untergebracht. Zunächst erhielten die Studierenden eine kurze Instruktion zur Funktionsweise der Übung sowie einige Empfehlungen zu ihrer sinnvollen Nutzung. Die Angaben enthielten auch einige pädagogische Tipps, z.B. sich die Erklärungen der Aufgabenbesprechung verständlich zu machen. Es wurde stets darauf hingewiesen, wissenschaftliche Studien hätten Belege vorweisen können, dass die adäquate Bearbeitung der Aufgaben Wissen stabilisiert und das Lernen fördert.

Nachfolgend waren die einzelnen Aufgaben hintereinander aufgelistet und konnten in beliebiger Reihenfolge in Angriff genommen werden. Als Aufgabentypen kamen fast ausschließlich spezielle Zuordnungs- oder True/False-Aufgaben zum Einsatz. Eine typische Zuordnungsaufgabe kann hierbei auch als eine Serie mehrerer, meist 4 bis 6 MC-Aufgaben gedeutet werden (siehe Aufgabenbeispiele unten). Ähnliches gilt für eine True/False Aufgabe, die in der Regel als Verbund mehrerer True/False-Antworten konzipiert war. Die erste Übung (Skalen) umfasste 7 Aufgaben, die insgesamt 40 Antworten verlangten. Die 8 Aufgaben der zweiten Übung (Gütekriterien) erforderten insgesamt 35 Antworten.

Bei allen Übungen musste jede Aufgabenbearbeitung eigens bestätigt werden. Es gab aber keinerlei Notwendigkeit, dies sofort zu tun, etwa bevor weitere Aufgaben in Augenschein genommen wurden. In einem Quiz musste der Studierende jedoch irgendwann im Verlauf der Übung auf den Button "Aufgabe bestätigen" klicken. Damit war diese Aufgabe beantwortet und das entsprechende Aufgabenergebnis somit festgelegt. Echte, die Aufgabenbewertung beeinflussende Aufgabenrevisionen, konnten somit nur vor dem Anklicken des Buttons vorgenommen werden. Unter jeder Musterlösung befand sich ein Button „Musterlösung bestätigen", den der Studierende ebenfalls im Verlauf der Übung anklicken musste, wodurch sich der Button zum Button „ok" veränderte.

Beim Versuch, durch Anklicken des Buttons „Übung beenden" die Seite zu verlassen, prüfte das Programm zunächst die Vollständigkeit der Aufgabenbearbeitungsbestätigungen und mahnte bei Nichterfüllung dieser Bedingung die noch zu bearbeitenden Aufgaben an. Die Übung konnte erst dann beendet werden, nachdem alle Aufgaben bearbeitet worden waren. Anschließend erschien ein Dialogfeld, in dem der Studierende sein Wissen in der betreffenden Übung nach einer Woche einschätzen sollte. (Beispiel für die **J**udgement **O**f **L**earning-Messung: „"Wie viel Prozent der Aufgaben des Quiz würden Sie vermutlich in einer Woche richtig beantworten? [Zahl zwischen 0 und 100 eingeben]").

Die speziellen experimentellen Bedingungen

Q_FAI: Quiz mit Feedback after item:

Nach der Bestätigung der Aufgabenbearbeitung einer Aufgabe erhielt der Studierende in einem speziellen Fenster unmittelbare Rückmeldung zu dieser Aufgabe hinsichtlich

- Knowledge of response (KOR: welche Antworten er richtig oder falsch beantwortet hatte)
- Knowledge of correct response (KCR: die korrekten Antworten)
- Elaboriertes Feedback (EL: nähere Erklärungen zur Aufgabenlösung und weitere Informationen)

KOR und KCR wurden obligatorisch, EL meistens, aber nur dann, rückgemeldet, wenn die Vermutung vorlag, viele Studierende könnten sich alleine die korrekte Antworten nicht hinreichend erklären. Zusätzlich hatte der Studierende erst nach der Aufgabenbestätigung die Möglichkeit, durch Anklicken des Buttons „korrekte Lösung?" die korrekte Aufgabenbearbeitung zu erzwingen und damit eine übersichtliche Musterlösung einzusehen. Grundsätzlich bestand die Möglichkeit, die Aufgaben in beliebiger Reihenfolge und beliebig oft zu bearbeiten und so die speziellen Rückmeldungen einzufordern. Unmittelbar nach der Quizbeendigung folgte die JOL-Messung und erst nachdem der Studierende seine Schätzung abgegeben hatte, erhielt er eine Rückmeldung zum Prozentsatz seiner korrekten Lösungen im Quiz (= Knowledge of performance).

Q_EOTF: Quiz mit End of Test Feedback

Die Testbedingungen bei Q_EOTF entsprachen exakt denen des Q_FAI. Nach Bestätigung einer Aufgabenbearbeitung folgte aber keinerlei Rückmeldung, sondern eine Deaktivierung der Aufgabenbestätigung für diese Aufgabe. Erst nach der Bestätigung aller Aufgabenbearbeitungen konnte der Studierende den Button „Quizbearbeitung beenden" erfolgreich aktivieren. Dann wurde ihm mitgeteilt, nun könne er durch eine erneute, jetzt wieder mögliche Aufgabenbestätigung die Auswertung seiner Bearbeitung sowie diverse Rückmeldungen zur Aufgabe einsehen und solle diese Chance nutzen. Zugleich initiierte dieses Ereignis die Zeitmessung für die Feedbackphase. Durch Anklicken des Buttons „Quiz endgültig beenden" war die Übung und die Zeitmessung der Feedbackphase beendet. Es schloss sich dann die JOL-Messung für diese Bedingung an. Erst danach wurde dem Studierenden der Prozentsatz seiner korrekten Lösungen im Quiz mitgeteilt (= Feedback of performance).

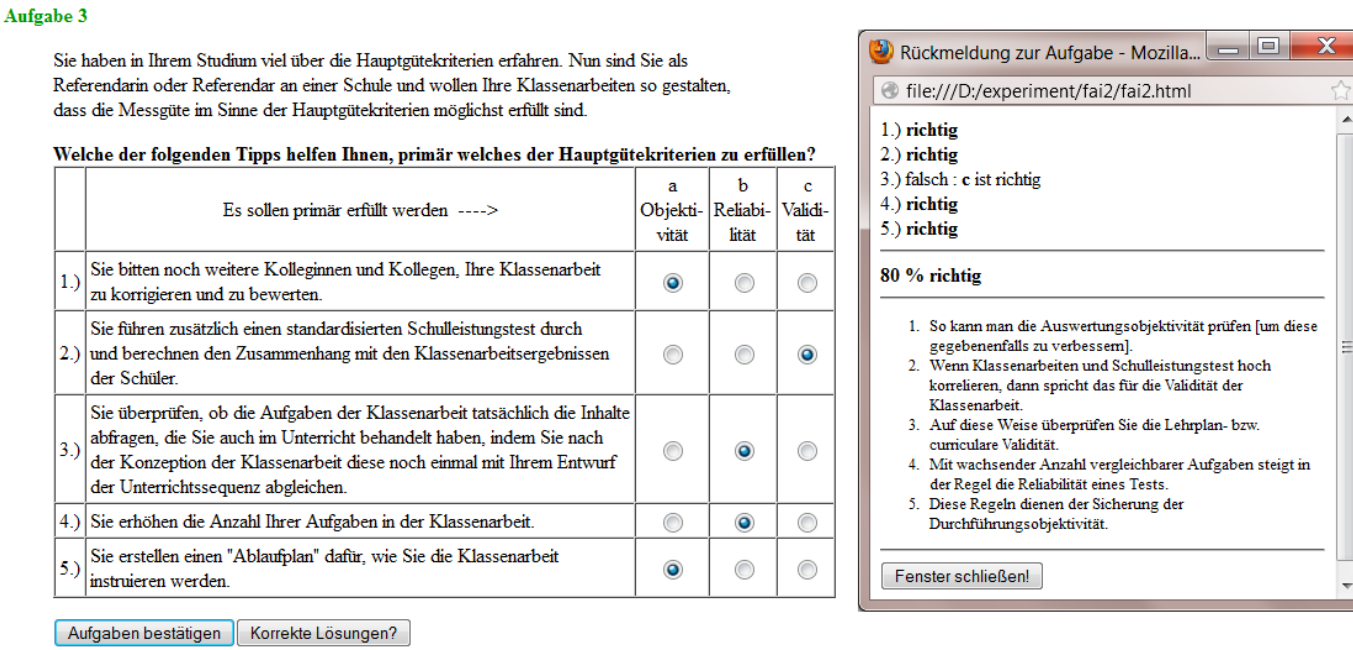
MU: Musterlösung

Die Musterlösung einer Aufgabe entspricht der direkten Darbietung einer korrekten Quizbearbeitung einschließlich des möglichen elaborierten Feedbacks. Die Studierenden sahen folglich eine HTML-Seite mit allen korrekt beantworteten Aufgaben einschließlich der Aufgabenbesprechungen. Ähnlich wie bei Q_FAI konnte die Übung erst nach Bestätigung aller Musterlösungen beendet werden. Unmittelbar danach folgte die JOL-Messung.

Aufgabenbeispiele

Nachfolgende Abbildungen 1a bis 1d demonstrieren zwei Übungsaufgaben sowie ihre dazu gehörigen Lernerfolgsaufgaben. Die experimentelle Variation bezieht sich nur auf die Übungsaufgaben, während alle Studierenden dieselben Lernerfolgsaufgaben bearbeiteten.

Abbildung 1a: Aufgabenbeispiel: Quiz mit Feedback after Item (Q_FAI)
Aufgabe 3 aus Übung 2



Aufgabe 3

Sie haben in Ihrem Studium viel über die Hauptgütekriterien erfahren. Nun sind Sie als Referendarin oder Referendar an einer Schule und wollen Ihre Klassenarbeiten so gestalten, dass die Messgüte im Sinne der Hauptgütekriterien möglichst erfüllt sind.

Welche der folgenden Tipps helfen Ihnen, primär welches der Hauptgütekriterien zu erfüllen?

	Es sollen primär erfüllt werden ---->	a Objekti-vität	b Reliabi-lität	c Validi-tät
1.)	Sie bitten noch weitere Kolleginnen und Kollegen, Ihre Klassenarbeit zu korrigieren und zu bewerten.	◉	○	○
2.)	Sie führen zusätzlich einen standardisierten Schulleistungstest durch und berechnen den Zusammenhang mit den Klassenarbeitsergebnissen der Schüler.	○	○	◉
3.)	Sie überprüfen, ob die Aufgaben der Klassenarbeit tatsächlich die Inhalte abfragen, die Sie auch im Unterricht behandelt haben, indem Sie nach der Konzeption der Klassenarbeit diese noch einmal mit Ihrem Entwurf der Unterrichtssequenz abgleichen.	○	◉	○
4.)	Sie erhöhen die Anzahl Ihrer Aufgaben in der Klassenarbeit.	○	◉	○
5.)	Sie erstellen einen "Ablaufplan" dafür, wie Sie die Klassenarbeit instruieren werden.	◉	○	○

Aufgaben bestätigen | Korrekte Lösungen?

Rückmeldung zur Aufgabe - Mozilla...

file:///D:/experiment/fai2/fai2.html

1.) **richtig**
2.) **richtig**
3.) falsch : **c** ist richtig
4.) **richtig**
5.) **richtig**

80 % richtig

1. So kann man die Auswertungsobjektivität prüfen [um diese gegebenenfalls zu verbessern].
2. Wenn Klassenarbeiten und Schulleistungstest hoch korrelieren, dann spricht das für die Validität der Klassenarbeit.
3. Auf diese Weise überprüfen Sie die Lehrplan- bzw. curriculare Validität.
4. Mit wachsender Anzahl vergleichbarer Aufgaben steigt in der Regel die Reliabilität eines Tests.
5. Diese Regeln dienen der Sicherung der Durchführungsobjektivität.

Fenster schließen!

Der Studierende hat die Aufgabe bearbeitet, auf „Aufgaben bestätigen" geklickt und so das Rückmeldefenster zum Vorschein gebracht. Würde er jetzt auf „Korrekte Lösungen?" klicken, so sähe er nur die richtigen Lösungen d.h. der Fehler bei 3.3. würde automatisch korrigiert werden. Unter Bedingung **Quiz mit End of Test Feedback (Q_EOTF)** ist der Button „Korrekte Lösungen" in der Testphase ständig deaktiviert und beim Anklicken auf „Aufgaben bestätigen" erscheint kein Rückmeldefenster. Stattdessen wird die Aufgabe intern bewertet und der Bestätigungsbutton deaktiviert. In der Feedbackphase nach Bearbeitung aller Aufgaben sieht der Studierende seine Antworten und kann das Feedback analog der Bedingung Q_FAI nun einfordern. Unter der Bedingung **Musterlösung** würde man die korrekt beantwortete Aufgabe und darunter den Text zur Aufgabenbesprechung [siehe Rückmeldefenster] sehen.

Abbildung 1b zeigt die Lernerfolgsaufgabe für die Übungsaufgabe in Abbildung 1a. Beide Aufgaben beinhalten die gleiche übergeordnete Frage, aber unterscheiden sich in etlichen Aspekten im Hinblick auf die speziellen Zuordnungen (Tipps). Ein Vergleich von Abbildung 1c und 1d lässt erkennen, dass die Übungsaufgabe und die ihr zugeordnete Lernerfolgsaufgabe unterschiedliche Beispiele umfassen.

Abbildung 1b: Aufgabe im Lernerfolgstest zur Aufgabe 3 der Übung 2
[= ist die Nummer 10 im Lernerfolgstest]

Aufgabe 10

Sie haben in Ihrem Studium viel über die Hauptgütekriterien erfahren. Nun sind Sie als Referendarin oder Referendar an einer Schule und wollen Ihre Klassenarbeiten so gestalten, dass die Hauptgütekriterien möglichst erfüllt sind.
Welche der folgenden Tipps helfen Ihnen, primär welches der Hauptgütekriterien zu erfüllen

	Es sollen primär erfüllt werden ---->	a Objekti-vität	b Reliabi-lität	c Validi-tät
1.)	Sie entwickeln dezidierte Richtlinien für die Aufgabenbewertung, aus denen hervorgeht,welche möglichen Antworten richtig bzw. falsch sind.	○	○	○
2.)	Sie überprüfen, ob die Aufgaben der Klassenarbeit tatsächlich die Leistungen abfragen, die im Lehrplan stehen und die sie zugleich auch im Unterricht behandelt haben.	○	○	○
3.)	Sie setzen einen standardisierten Test ein, der sehr ähnliche Lehrziele wie ihre Klassenarbeit zu erheben beansprucht, und berechnen den Zusammen-hang.	○	○	○
4.)	Sie verdoppeln die Anzahl Ihrer Aufgaben in der Klassenarbeit.	○	○	○
5.)	Sie verändern das Aufgabenformat für einige Fragen, so dass die Korrektheit der Antwort besser festgestellt werden kann.	○	○	○

Abbildung 1c: Musterlösung aus Übung 1

Musterlösung 1

Geben Sie das höchstmögliche Skalenniveau an, welches die Skala annehmen kann.

	Skala	a nominal	b ordinal	c intervall	d rational
1.)	Haarfarbe	◉	○	○	○
2.)	Temperatur in Celsius	○	○	◉	○
3.)	Intelligenzquotient	○	○	◉	○
4.)	Körpergröße in cm	○	○	○	◉
5.)	Fehleranzahl im Diktat	○	○	○	◉
6.)	Konfession	◉	○	○	○
7.)	Hochschulranking	○	◉	○	○
8.)	Schulabschluss	○	◉	○	○

Versuchen Sie die korrekten Antworten näher zu begründen.
Bearbeitung der Musterlöung bestätigen !

Die Aufgabenstellung des Übungsbeispiels in Abbildung 1c sah keine elaborierte Rückmeldung vor. Unter den Quizbedingungen müsste der Studierende die Aufgabe selbst beantworten und würde als Feedback eine Auswertung seiner Ergebnisse analog

obigem Beispiel in Abbildung 1a sowie die Anregung „Versuchen Sie die korrekten Antworten näher zu begründen" erhalten.

Abbildung 1d: Lernerfolgsaufgabe zur Aufgabe 1 in der Übung 1

Aufgabe 1

Geben Sie das höchstmögliche Skalenniveau an, welches die Skala annehmen kann.

	Skala	a nominal	b ordinal	c intervall	d verhältnis
1.)	Nationalität	○	○	○	○
2.)	Temperatur in Fahrenheit	○	○	○	○
3.)	Hantelgewicht in Kilogramm	○	○	○	○
4.)	Intelligenztest	○	○	○	○
5.)	Anzahl der Tippfehler	○	○	○	○
6.)	Autofarbe	○	○	○	○
7.)	höchster Bildungsabschluss	○	○	○	○
8.)	Beliebtheit deutscher Universitäten	○	○	○	○

Den genauen Ablauf der Übungen unter Q_FAI, Q_EOTF und Musterlösung findet man exemplarisch unter der URL:
http://bildungswissenschaften.uni-saarland.de/personal/jacobs/artikel/treatments/fai_eotf_muster/index.html

Testergebnisse der Übungsquiz

Um die Zuverlässigkeit der Befunde zu stärken, wurden die Testergebnisse und Reliabilitäten der Quiz auf der Basis aller an den Quiz teilnehmenden Studierenden ermittelt, also auch der ca. 15% Studierender, die den Lernerfolgstest später nicht bearbeitet hatten. Bei der Auswertung lag eine milde Zufallskorrektur zugrunde, welche vornehmlich die hohe Ratewahrscheinlichkeit bei den True/False-Aufgaben korrigieren sollte. Spearman Rangkorrelationen zwischen dem Prozentsatz der korrekten Lösungen und den Quizbearbeitungszeiten weisen in die Richtung, dass bessere Quizleistungen mit einer etwas längeren Bearbeitungszeit einhergehen (Quiz 1: rs = .22 (p = .015, nicht signifikant; Quiz 2: rs=.26, p= 0.007). Die Quizergebnisse korrelieren außerdem konsistent signifikant in erwarteter Richtung mit dem Abiturnotendurchschnitt (Quiz 1 r = -.29 bzw. Quiz 2: r = -.23) sowie mit den Angaben der Studierenden, sie hätten die Folie zur entsprechenden Vorlesung durchgearbeitet (Quiz 1: r = -.23 bzw. Quiz 2: r = -.29).

Wie aus Tabelle 3 hervorgeht, erzielten die Studierenden in den zwei Quiz Erfolgsquoten von jeweils ca. 60%, wobei nach t-Test für unabhängige Stichproben stets keinerlei Leistungsunterschiede zwischen den beiden Quizvarianten festzustellen waren. Die Ermittlung der Reliabilität nach Cronbachs α basierte daher auf der Zusammenfassung der Daten beider Quizvarianten.

Tabelle 3: Prozentsatz korrekter Lösungen in den Quiz zur Übung 1 (Skalen) und Übung 2 (Testgütekriterien)

	α	M	s	N	t	pz
Q_FAI_1		59.4	22.0	59		
	.78				-0.93	.355
Q_EOTF_1		63.1	20.6	60		
Q_FAI_2		58.1	20.4	54		
	.69				-0.07	.95
Q_EOTF_2		58.4	18.1	51		

Die Reliabilitäten von $\alpha = .78$ und .69 erreichen eine für Gruppenanalysen hinreichende Höhe. Wegen der Zufallszuteilung der Probanden reflektieren die Ergebnisse der Quiz-Bedingungen ein repräsentatives Bild der Lernleistung zum gegebenen Zeitpunkt für alle Gruppen. Die mäßigen Erfolgsquoten legen die Vermutung nahe, die Rückmeldungen der Quizvarianten seien potenziell lernförderlich einzuschätzen, da sie bei ca. 40 % Fehlern noch hinreichende Korrekturmöglichkeiten bieten. Denn der wesentliche Lerneffekt des Feedbacks liegt nicht in der Bestätigung korrekter, sondern der Korrektur falscher Aufgaben.

Ergebnisse

Bearbeitungszeiten für die experimentellen Übungen

Als Bearbeitungszeit für eine Übung wurde die Zeitspanne zwischen dem Erscheinen und Verschwinden der Übungsseite auf dem Bildschirm gemessen. Obgleich die Studierenden angewiesen wurden, sich nur der Übung zu widmen und das gesamte Programm ohne Unterbrechungen durchzuziehen, bleibt offen, ob sie sich daran gehalten haben. Die so erfasste Zeit kann daher nur bedingt als aktive Lernzeit betrachtet werden, da keinerlei externe Kontrolle vorliegt. Im Datensatz lassen sich sowohl ultrakurze wie überlange Zeiten identifizieren. Um die Ausreißerproblematik nach oben und unten zu entschärfen, erschien es ratsam, als zentralen Wert der Übungszeit einer experimentellen Bedingung den jeweiligen Median heranzuziehen und alle statistischen Vergleiche mit nicht parametrischen Verfahren durchzuführen. Wegen der Zufallszuteilung zu den experimentellen Bedingungen ist davon auszugehen, das zumindest die Treatmentunterschiede im Median der Bearbeitungszeiten die Unterschiede der tatsächlich investierten oder zumindest die in der Universitätspraxis eingesetzten Übungszeiten hinreichend approximieren.

Wie die statistischen Analysen belegen und aus Tabelle 4 augenscheinlich hervorgeht, lassen sich keinerlei Zeitunterschiede zwischen beiden Feedbackvarianten ermitteln. Bei EOTF setzt sich die Arbeitszeit aus reiner Testung und Feedbackphase zusammen, während bei FAI Testung und Rückmeldung meist im Wechsel miteinander vonstatten gehen. Im Gegensatz zu EOTF, wo Test- und Feedbackzeit voneinander separierbar sind, konnte diese Trennung unter FAI nicht vorgenommen werden. Die letztlich vergleichbaren Gesamtübungszeiten von FAI und EOTF deuten aber darauf hin, das Feedback sei von der Nutzungsdauer ähnlich einzuschätzen.

Theoretische Überlegungen sowie die bisherigen Erfahrungen ließen kürzere Bearbeitungszeiten für die Musterlösungen erwarten. Die Ergebnisse fallen allerdings nicht

überzeugend konsistent aus. Bemüht man den intraindividuellen Vergleich, so bearbeiteten die Studierenden ihr Quiz zwar stets länger als ihre Musterlösung. Allerdings trat dieser Effekt nur dann deutlich in Erscheinung, wenn erst das Quiz [hier Skalen] und dann die Musterlösung [Gütekriterien] bearbeitet wurden. Umgekehrt beträgt der Vorteil nur etwas mehr als eine Minute. Das bestätigten auch die hier angestrengten unabhängigen Vergleiche. Die erwarteten längeren Bearbeitungszeiten der Quiz lassen sich nur beim Übungsthema Gütekriterien (bzw. bei Übung 2) statistisch belegen, dort allerdings im Ausmaß hoher praktischer Bedeutsamkeit.

Tabelle 4: Bearbeitungszeiten der Übungen in Sekunden

Mediane der Zeiten in Sec		
	Übung 1	Übung 2
	Skalen	Testgüte
Q_FAI	**800**	**790**
Muster	**718**	**496**
z=	0.995	5.29
p=	0.320	<0.001
Q_EOTF	**778**	**850**
Muster	**673**	**423**
z=	0.458	5.26
p=	0,640	<.001
pro Gruppe	N=51-56	n= 47-58

Tabelle 5: Prozentualer Anteil der Musterlösungszeit an der Quizzeit
[(Zeit Musterlösung/Zeit Quiz)*100%]

	Übung 1	Übung 2
Q_FAI	**90**	**63**
Q_EOTF	**87**	**50**

Tabelle 5 fasst die Ergebnisse als Prozentsätze der Bearbeitungszeit für die Musterlösung gemessen an der Quizbearbeitungszeit zur selben Übung noch einmal zusammen. Der Mittelwert dieser Prozentsätze liegt mit ca. 75% im Rahmen der bisherigen Ergebnisse von Jacobs und bestätigt somit die Zeitersparnis durch die Musterlösung. Das Ausmaß des Zeitvorteils der Musterlösung hängt aber offensichtlich auch von der Reihenfolge ab. Nach einer Testung erscheint die Bearbeitung einer Musterlösung womöglich nicht mehr so wichtig, so dass diese insbesondere unter Zeitdruck, der gegen Ende der Sitzung stärker zunimmt, eher an Interesse verliert als bei einer Testung, die zumindest formell einen höheren Aufforderungscharakter ausübt. Ähnliche Effekte könnten auch bei solchen Studien eingetreten sein, die in einem klassischen Wiederholungsdesign Test- und Musterlösungsitems an unterschiedlichen Positionen vorgaben.

Eine Häufigkeitsanalyse aller Bearbeitungszeiten offenbart einige unliebsame Unzulänglichkeiten pädagogischer Praxis und soll an dieser Stelle kurz skizziert werden. Ca. 15% aller Studierenden nahmen sich für eine Übung mit Musterlösungen höchstens 3 Minuten Zeit, was kaum ausreicht, den vorliegenden Text der Übung überhaupt aufmerksam lesen zu können. Bei einigen Studierenden, deren Anteil glücklicherweise nur bei 5% liegt, kommt der begründete Verdacht auf, sie hätten sich ohne jede ernsthafte Kenntnisnahme durch die Musterlösungen geklickt, um die Übung beenden zu können. Ultrakurze Bearbeitungszeiten für Quiz waren sehr selten. Nur 2% beendeten ein Quiz innerhalb von 3 Minuten. Unter der Feedbackvariante Feedback after Test war es möglich, Test- von Feedbackzeit zu trennen. 27% aller Studierenden unter der Bedingung EOFT schauten sich das Feedback weniger als eine Minute an, der Median liegt knapp unter 2 Minuten. Angesichts der Erfolgsquoten von ca. 60% in beiden Quiz reichen diese Zeiten in vielen Fällen kaum aus, ein hinreichendes Verständnis aufzubauen. Bei diesen Ergebnissen ist jedoch zu beachten, dass die Studierenden die Übung weniger aus eigenem Antrieb heraus anstrebten, sondern für viele überraschend zu einer solchen Übungsbearbeitung bewegt wurden, die für manche in der dafür veranschlagten Zeit nicht in ihr Planungskonzept passte. Mit derartigen Problemen wird man in der pädagogischen Praxis häufiger konfrontiert, insbesondere dann, wenn man auf leistungsabhängige Konsequenzen verzichtet.

Ergebnisse im Lernerfolgstest

Die Studierenden bearbeiteten den Lernerfolgstest unter kontrollierten Bedingungen in der Vorlesung im Papier und Bleistift-Format. Ca. 5 bis 10 Studierende pro Bedingung traten nicht mehr zum Lernerfolgstest an oder konnten nicht zugeordnet werden, was einen Ausfall von insgesamt ca. 15% verursachte, der angesichts seines geringen Ausmaßes und der Randomisierung unproblematisch erscheint. In der betreffenden Vorlesung bearbeiteten aber auch viele Studierende den Lernerfolgstest, die zuvor an keiner Maßnahme (auch nicht als ausgewählte Kontrollgruppe) teilgenommen hatten. Diese fungieren hier als zusätzliche Kontrollgruppe KG'.

Der Zeitabstand zwischen Übung und Lernerfolgstest schwankte je nach Übungsantritt des Studierenden zwischen 12 und 7 Tage, betrug folglich mindestens eine Woche. Insgesamt umfasste der Lernerfolgstest 15 Aufgaben, die als Parallelitems zu den Aufgaben der entsprechenden Quiz konstruiert wurden und in die Subtests Lernerfolg 1 und Lernerfolg 2 getrennt werden. Auf eine Zufallskorrektur wurde diesmal verzichtet und als Messvariable diente jeweils der Prozentsatz korrekter Lösungen gemessen an allen Antworten. Lernerfolg 1 (7 Aufgaben mit insgesamt 40 Antworten) und Lernerfolg 2 (8 Aufgaben mit insgesamt 38 Antworten) erzielten jeweils eine Reliabilität von $\alpha = .80$ (N = 348). Mit Hilfe des Lernerfolgs kann die Parallelretestreliabilität der beiden Quiz ermittelt werden. Die Korrelation zwischen Quiz 1 und Lernerfolg 1 beträgt $r = .72$ (N = 104) und die vom Quiz 2 mit Lernerfolg 2 $r = .58$ (N = 88). Beide Lernerfolgsmaße korrelieren zu .61 miteinander.

Tabelle 6 stellt die Lernerfolgsergebnisse für alle experimentellen Gruppen detailliert dar und macht kenntlich, welche Übungsmethode jeweils angewandt wurde, deren Auswirkung im Lernerfolgstest ermittelt werden sollte.

Tabelle 6: Prozentsatz korrekter Antworten im Lernerfolgstest

		Übung 1			Übung 2		
	N		M	s	M	s	
Gruppe 1	54	Q_FAI	67.7	15.7	66.6	14.1	Muster
Gruppe 2	44	Muster	65.8	12.9	69.7	15.3	Q_FAI
Gruppe 3	50	Q_EOTF	65.4	14.3	71.9	13.4	Muster
Gruppe 4	44	Muster	66.3	11.4	66.9	15.3	Q_EOTF
Gruppe 5	52	KG	61.9	13.4	63.0	13.9	KG
Rest	104	KG'	60.5	13.5	58.7	16.0	KG’

Aus der Tabelle 6 kann man z.B. ablesen, welche Lernerfolgsergebnisse die 44 Studierenden in Gruppe 2 erzielten, die in der Übung 1 die Musterlösung und in Übung 2 das Quiz mit FAI bearbeiteten. Beim direkten Vergleich der beiden Feedbackvarianten zu den jeweiligen Übungen findet man konsistent sehr ähnliche Lernerfolgsmaße. Eine statistische Prüfung mit t-Test für unabhängige Stichproben ergab weder für Übung 1 ($t(102) = 0.776$, $p = 0.44$) noch für Übung 2 ($t(86)= 0.84$, $p = 0.4$) signifikante Unterschiede zwischen beiden Feedbackvarianten, weswegen im Folgenden auf eine weitere Differenzierung nach Feedbacktypen verzichtet wird.

Tabelle 7 basiert auf einigen Datenzusammenfassungen und zeigt die Lernerfolgsergebnisse für alle Quiz- und Musterlösungsbedingungen der jeweiligen Übung. Die erste Datenzeile bezieht sich z.B. auf die Lernerfolgsergebnisse derjenigen Studierenden, die entweder in Übung 1 oder in Übung 2 ein Quiz bearbeiteten. Zusätzlich wurden die Lernerfolgsergebnisse der beiden Kontrollgruppen einbezogen.

Tabelle 7: Zusammenfassende Ergebnisse im Lernerfolgstest

	Lernerfolg 1 für Übung 1		Lernerfolg 2 für Übung 2	
	M	s	M	s
Quiz	66.6	15.0	68.3	15.3
Muster	66.0	12.1	69.1	13.9
KG	61.9	13.4	63.0	13.9
KG'	60.5	13.5	58.7	16.0

Die Anzahl der Probanden bei den Quiz und den Musterlösungen beträgt je nach Übung 88 oder 104. Die Anzahl der Probanden für die Kontrollgruppe KG beträgt 52, die der KG' 104.

Tabelle 8 zeigt die Ergebnisse der statistischen Analyse mittels LSD-multiplem Mittelwertsvergleich

Tabelle 8: Mittelwertsunterschiede im Lernerfolgstest und Effektstärke (Cohens d)

	Übung 1		**Übung 2**	
Übung 1:	**p**	**d**	**p**	**d**
Quiz vs Muster	**.775**		**.702**	
Quiz vs KG	**.022**	**.33**	**.022**	**.36**
Quiz vs. KG'	.0005	.43	.008	.61
Muster vs. KG	**.043**	**.32**	**<.0005**	**.46**
Muster vs. KG'	.0025	.43	<.0005	.70

Anmerkung: Beim Vergleich Quiz vs. Muster wurde zweiseitig, ansonsten jeweils einseitig getestet. Cohens d wurde nur bei signifikanten Ergebnissen berechnet.

Wie aus den Tabellen 7 und 8 hervorgeht, lassen sich weder für Übung 1 noch für Übung 2 signifikante Unterschiede im Lernerfolgstest zwischen den Quiz und den Musterlösungen nachweisen. Die hoch vergleichbaren Lernerfolgsmaße beider Bedingungen bestätigen somit die Ausgangshypothese, Musterlösungen würden ähnliche Lerneffekte bewirken wie Testen mit Feedback. Den Ergebnissen zufolge spricht die in Übung 2 realisierte deutlich geringere Übungszeit der Musterlösung gegenüber den Quiz (siehe Tabelle 4 und 5) weniger für eine oberflächliche Bearbeitung, sondern eher für eine effizientere Informationsnutzung.

Musterlösungen sowie Testen mit Feedback sollten zumindest eine längerfristige Lernwirkung hinterlassen, die erkennbar höher ausfällt als die Bearbeitung der vier kurzen Online-Konzentrationstests der Kontrollgruppe, da diese Beschäftigung gar nichts mit den Lehrzielen der Übung zu tun hatte. Wie die statistischen Analysen aufzeigen, ließ sich diese Vermutung für jede Übung bestätigen, wenngleich das Ausmaß des Effektes konsistent im niedrigen Effektstärkebereich angesiedelt ist. Bei Zusammenfassung aller Treatmentgruppen auf den gesamten Lernerfolg fällt der Vorteil der Onlineübung gegenüber der Kontrollgruppe erwartungsgemäß statistisch ganz eindeutig aus ($p = 0.006$). D.h., die angebotene Onlineübung, die ja stets ein Quiz und eine Musterlösungsvariante beinhaltete, bewirkte im Mittel einen statistisch gut abgesicherten, kleinen Lerngewinn. Der deutlich günstigere Vergleich der Treatmentgruppen mit der KG' ist in seiner Aussagekraft jedoch beschränkt, da es sich um eine vorgegebene und nicht durch Randomisierung kontrollierte Gruppe handelt.

Wertschätzung der Übungsmethoden

Aufgrund früherer Untersuchungen von Jacobs (2010, 2011) war mit sehr hoher Evidenz zu erwarten, Studierende würden einem Quiz gegenüber der Musterlösung eine deutlich höhere Wertschätzung entgegen bringen. Zur Messung dieser Einschätzung wurden einige Items von Jacobs übernommen und leicht modifiziert. Sie beziehen sich zum einen auf eine Bewertung der Übungsmethode anhand einer 6-stufigen Likertskala hinsichtlich ihrer Lernwirksamkeit, pädagogischen Qualität und Interessensweckung. Zum anderen verlangen sie eine Angabe darüber, wie stark der Studierende selbst die Übungsmethode präferieren oder weiter empfehlen würde. Der Fragebogen umfasst 5 Items, die im Gegensatz zur Forced-choice-Methode bei Jacobs, getrennt für das Quiz und die Musterlösung zu beantworten waren. Für das Quiz erzielte die Wertschätzung eine Reliabilität von $\alpha = .93$ und für die Musterlösung ein $\alpha = .95$. Außerdem bewerteten die Studierenden die subjektive Qualität der Übungsmethoden hinsichtlich des pädagogischen Nutzens auf einer 11 Punkteskala, deren Enden mit „0 = unbrauchbar" bzw. „10 = hervorragend" verbal verankert waren. Zudem sollten sie für jede Übungsform eine Note im Bereich von 1 bis 6 vergeben. Die drei Wertschät-

zungsvariablen für die Quiz sowie die für Musterlösungen korrelieren dem Betrage nach zwischen .66 und .83 miteinander. Hohe Bewertungen der Quiz gehen einher mit signifikant besseren Quizleistungen (Korrelationen im Bereich eines Betrages von .22 bis .36) und einer längeren Nutzungsdauer der Quiz bzw. der Feedbackbearbeitungszeit (Rangkorrelationen im Bereich eines Betrages von .15 bis .21 bzw. .28 bis .38).

Wie aus den Tabellen 9a und b hervorgeht, sprechen alle Ergebnisse eindeutig für eine höhere Wertschätzung der Quiz gegenüber den Musterlösungen. Tabelle 9a zeigt für jedes Übungsthema, differenziert nach der Feedbackversion des Quiz, stets eine bessere Bewertung der Studierenden für das Quiz im Vergleich zur Einschätzung derjenigen Studierenden, welche zum gleichen Übungsthema die Musterlösung bearbeiteten. Alle unabhängigen Vergleiche basieren durch die Randomisierung auf experimentellem Niveau und erzielten mindestens mittlere Effektstärken zugunsten der Quiz.

Tabelle 9a: Wertschätzung für jedes Übungsthema und jede Feedbackversion (N pro Vergleich = 47-58)

unabhängige Vergleiche

	Übung 1		Übung 2	
	Skalen		Gütekriterien	
	M	s	M	s
Q_FAI	24.9	4.8	25.7	3.6
MU	21.3	5.9	19.9	5.8
d	**.67**		**.91**	
Q_EOTF	24.7	4.6	24.1	4.2
MU	21.0	6.3	21.5	5.9
d	**.67**		**.51**	

Tabelle 9b: Bewertung von Quiz und Musterlösung für alle TeilnehmerInnen (t-Test für abhängige Stichproben, N=224)

		M	**s**	**t**	**d**
Wertschätzung	Quiz	25.1	4.3	8.7	**.77**
	Musterlösung	21.1	5.9		
Päd. Nutzen	Quiz	8.1	1.4	10.1	**.98**
	Musterlösung	6.4	2.1		
Note	Quiz	1.9	0.8	-9.2	**.84**
	Musterlösung	2.7	1.1		

Tabelle 9b prüft, wie die Studierenden ihr Quiz gegenüber ihrer Musterlösung bewerteten und auch bei dieser Sichtweise auf der Basis eines Wiederholungsdesigns ergibt sich eine deutlich bessere Wertschätzung für die Quiz. In die gleiche Richtung weisen die Ergebnisse der Einzelskalen zum pädagogischen Nutzen und der Notenbewertung.

Wie aus Tabelle 9a weiterhin ersichtlich ist, sind keinerlei Unterschiede zwischen den Feedbackvarianten der Quiz zu erkennen, was im Übrigen auch für die Einschätzung

des pädagogischen Nutzens und die Note zutrifft. Für die subjektive Wertschätzung scheint es somit keine Rolle zu spielen, ob die Rückmeldung direkt nach jeder Aufgabenbearbeitung oder erst am Ende eines Tests kommt. Allerdings fehlte den Studierenden für diesen Vergleich im Gegensatz zum Vergleich mit der Musterlösung die eigene Erfahrung, weil sie entweder ein Quiz mit FAI oder ein Quiz mit EOTF bearbeiteten.

Wahrgenommene Beanspruchung der Übungsmethoden

Mit 5 Items sollte abgeprüft werden, ob das Testen im Vergleich zur Musterlösung aus der Sicht der Studierenden eine höhere Beanspruchung im Sinne von mehr Angst, Stress, Bedenken, Anstrengung oder Schwierigkeit verlangte. Hierbei waren die meisten Items so formuliert, auch kleine Beeinträchtigungen erfassen zu können (Itembeispiel: „Ich fand die Aufgabenbearbeitung teilweise stressig."). Die entsprechenden Items mussten getrennt für das Testen wie für die Musterlösung beantwortet werden und wurden zu einem jeweiligen Fragebogen „Beanspruchung" zusammengefasst, dessen Reliabilitäten nach Cronbach für das Testen $\alpha = .78$ und für die Musterlösung $\alpha = .75$ ergaben. Die subjektive Beanspruchung der Quiz korreliert erwartungsgemäß signifikant negativ mit dem Prozentsatz der korrekten Lösungen im jeweiligen Quiz (r = -.39 sowie r = -.44; N jeweils um 100) sowie ebenso mit der Einschätzung des Prozentsatzes der korrekten Lösungen in einer Woche (JOL, r = -.19 bzw. r = -.40; N = 83-98). Je schwächer das objektive Quizergebnis und die etwas langfristigere Einschätzung der Quizleistung, desto höher wird die Beanspruchung des Testens mit Feedback erlebt, allerdings in ähnlicher Größenordnung auch die Beanspruchung der Musterlösung beim anderen Übungsthema. Da von den meisten Studierenden Ergebnisse von Prüfungsangst- bzw. Prüfungsängstlichkeitsfragebögen vorlagen (SPA, FT, TAI), die ca. zwei Wochen vor dem Experiment erhoben wurden, lag es nahe, den erwarteten positiven Zusammenhang dieser Maße mit der Beanspruchung zu prüfen. Alle eingesetzten Testverfahren zur Prüfungsangst bzw. -ängstlichkeit korrelieren signifikant positiv mit der aktuellen Beanspruchung der Quiz in einem Bereich von .29 bis .37 und der Musterlösung im einem Bereich von .25 bis .31 (N jeweils ca. 200).

Hier geht es primär um die Frage, welchen Einfluss die Übungsmethoden auf die Beanspruchung ausübten. Im Gegensatz zu den Einschätzungen und Persönlichkeitsmerkmalen kann diese Frage experimentell geprüft werden. Wie aus Tabelle 10 hervorgeht, fiel die Beanspruchung beim Testen etwas höher aus als bei der Musterlösung. Eine Analyse der Items ergab, dass der Unterschied in der Beanspruchung im Wesentlichen auf das Item Angst: „Die Übungsmethode löst gelegentlich Angst und Verunsicherung aus." zurück geht. Hinsichtlich der eingeschätzten Anstrengung, die insgesamt unter dem Skalenmittelpunkt lag sowie der Aufgabenschwierigkeit, die in ihrer Gesamtheit als mittelschwer eingeschätzt wurde, ließen sich keine nennenswerten Unterschiede zwischen Testen und Musterlösung feststellen.

Abbildung 10: Unterschiede zwischen Testen und Musterlösung (t-Test für abhängige Stichproben (N = 224)

Fragebogen/Item		M	s	t	d
Beanspruchung	Quiz	15.7	4.7	4.5	**.25**
	Musterlösung	14.5	4.6		
Item Angst	Quiz	3.1	1.5	6.2	**.47**
	Musterlösung	2.4	1.5		

Anmerkung: Item Angst: „Die Übungsmethode löst gelegentlich Angst und Verunsicherung aus."

Unterschiede in der Beanspruchung zwischen den Quizmethoden?

Es war vermutet worden, die Aufgabenbearbeitung könne eher durch eine unmittelbare Rückmeldung nach jeder Aufgabe als durch das verzögerte Feedback nach der gesamten Quizbearbeitung gestört werden. Denn unmittelbares Feedback enthielt stets auch eine Bewertung der Aufgabenbearbeitung im Sinne von "richtig/falsch". Diese Rückmeldung ließe sich auch als permanenter Hinweis auf Erfolg und Misserfolg deuten, was die Aufmerksamkeit auf das Selbst lenken und so Zweifel an der eigenen Leistungsfähigkeit begünstigen könnte, wodurch eher Angst und Verunsicherung resultiere. Es traten jedoch zwischen den Feedbackvarianten in beiden Übungen keinerlei signifikante Unterschiede in der Beanspruchung oder einer ihrer Items auf, so dass diese Hypothese eindeutig verworfen werden musste. Vermutlich wirkt die durch permanente Rückmeldungen aktivierte, potenzielle Erwartung von Erfolg oder Misserfolg nur in klassischen Prüfungssituationen hinreichend bedrohlich.

Insgesamt weisen die Ergebnisse zur Beanspruchung darauf hin, die Testung sei überwiegend im Sinne einer Übung und weniger als eine klassische Bewertungssituation gedeutet worden. Dies entsprach auch der pädagogischen Zielsetzung, eine unbenotete Übung anzubieten, die den Studierenden zudem volle Anonymität zusicherte. Natürlich kann man auch bei einer anonymen Testung selbst gesetzte Ziele verfehlen und sich eigenständig unter Stress setzen. Solche möglichen Reaktionen erwiesen sich allerdings im Mittel von geringer praktisch pädagogischer Relevanz. D.h. trotz der hier festgestellten geringfügig höheren Beanspruchung einer Testung gegenüber einer Musterlösung stellt das Übungsquiz mit unmittelbaren Rückmeldungen keine nennenswerte affektive Beeinträchtigung dar. Diese Interpretation steht im Einklang mit Befunden von Jacobs (2009, 2010), der unmittelbar vor unbenoteten Quiz mit FAI eher geringe aktuelle Angstmittelwerte feststellte, was im Übrigen auch die aktuellen Angstmessungen unmittelbar vor dem hier - ebenfalls unbenoteten - Lernerfolgstest bestätigten.

Judgement of learning (JOL)

Bisherige Untersuchungen weisen in die Richtung, reines Studieren, etwa das wiederholte Durchlesen eines Textes, begünstige Kompetenzillusionen und bewirke daher höhere Leistungseinschätzungen als das Testen, welches durch die Anforderungen und Rückmeldungen ein realistisches Bild der eigenen Leistungsfähigkeit widerspiegele. Zwar führten Lernerfolgskontrollen unmittelbar im Anschluss an die Übung auch meist zu besseren Ergebnissen reinen Studierens gegenüber einer Testung, womit höhere aktuelle Leistungseinschätzungen durch Studieren eine gewisse objektive Basis hätten, der Behaltensvorteil des Studierens verkehrt sich jedoch mit zunehmendem Retentionsintervall immer deutlicher in einen Nachteil gegenüber der Testung (z.B. Roediger & Karpicke, 2006b; Metaanalyse Rowland, 2014). Überhöhte Leistungsprognosen werden deshalb als pädagogisch bedenkliche Metakognitionen betrachtet, da sie die Betroffenen in einer ungerechtfertigen Sicherheit wiegen, welche sich negativ auf die Vorbereitung auswirken könnte. Um diese Hypothese zu testen und dabei eine etwas längerfristigere Leistungsprognose zu berücksichtigen, sollten die Studierenden nach Beendigung ihrer jeweiligen Übung einschätzen, wie viel Prozent der eben bearbeiteten Aufgaben sie in einer Woche richtig lösen würden. Unter Q_FAI erhielten die Studierenden zwingend eine objektive Bewertung nach jeder Aufgabenbearbeitung in Form von richtig oder falsch sowie bei umfangreichen Zuordnungsaufgaben zusätzlich den Anteil der korrekten Lösungen bei dieser Aufgabe, unter

Q_EOTF blieb es ihnen selbst überlassen, diese Information nach ihrer Quizbearbeitung anzufordern. Die Rückmeldung zum Prozentsatz der korrekten Lösungen folgte jedoch stets erst nach der JOL-Erhebung. Die objektiven Leistungsrückmeldungen vor dem JOL basierten also immer nur auf Itemniveau. Tabelle 11 verdeutlicht die Mittelwerte der Leistungsprognosen für die einzelnen Übungen.

Tabelle 11: Einschätzungen des **Prozentsatzes korrekter Lösungen** in einer Woche. Judgement of learning (N pro Vergleich = 47-58; nur Studierende, die Daten sowohl über Quiz wie Musterlösungen ablieferten)

unabhängige Vergleiche

	Zeitpunkt 1 Übung 1 Skalen	Zeitpunkt 2 Übung 2 Gütekriterien
FAI	73.4	64,6
MU	63,9	64,4
EOFT	76,4	66,4
MU	67,0	66,2

Während man bei Übung 2 keinerlei Unterschiede zwischen Quiz und Musterlösungen entdecken kann, fallen die Prognosen unter den Quizbedingungen bei Übung 1 eindeutig höher aus. Sie liegen im Mittel auch bedeutsam über den objektiven Ergebnissen der Quiz und deuten jedenfalls eher auf Kompetenzillusionen als die Prognosen unter der Bedingung Musterlösungen. Ein Vergleich des Abstands zwischen Schätzung und Quizergebnis ist aber insofern etwas problematisch, als Studierende bei Ihrer Einschätzung keine Zufallskorrektur durchführen. Für einen absoluten Vergleich eignet sich jedoch der Lernerfolgstests, da hier auf die Zufallskorrektur verzichtet und genauso wie bei Einschätzung der Studierenden exakt der Prozentsatz der korrekten Antworten zugrunde gelegt wurde.

Zunächst stellt sich die Frage, ob sich die Leistungsprognosen der Studierenden an den erbrachten objektiven Leistungsergebnissen orientieren, was nur in den Quizübungen überprüfbar ist. Des Weiteren kann für beide Übungsvarianten die Validität der studentischen Prognose in Form einer Korrelation ihrer Schätzung mit dem Lernerfolg ermittelt werden. Da Korrelationen aber keine Niveauunterschiede erfassen, wird als weiteres Genauigkeitsmaß der Betrag der Abweichung zwischen Schätzung nach der Übung und dem späteren Lernerfolg berechnet und anschließend die Genauigkeitsunterschiede zwischen Quiz und Musterlösung angestrengt.

Da sich weder in den JOL-Mittelwerten noch bzgl. der Höhe des Zusammenhangs mit dem objektiven Quiz- und Lernerfolgsergebnis Unterschiede zwischen den Quizversionen feststellen ließen, wurden beide Quizvarianten zusammengefasst. Unterschiede zwischen den Quizvarianten waren auch nicht erwartet worden, weil beide Feedbacktimingversionen letztlich die gleichen Rückmeldungen gewährten, lediglich zu verschiedenen Zeiten.

Tabelle 12: Korrelationen zwischen subjektiven Einschätzungen, Quizergebnis und Lernerfolgstest (N = 78-104)

		Ergebnis Quiz	Ergebnis Lernerfolg
Übung 1	JOL nach Quiz	.54	.50
	JOL nach Muster		.40
Übung 2	JOL nach Quiz	.53	.38
	JOL nach Muster		.45

Anmerkung: alle Koeffizienten sind auf dem Promilleniveau signifikant.

Wie aus Tabelle 12 hervorgeht, orientiert sich die Schätzung des Lernerfolgs in einer Woche an der tatsächlichen Leistung im Quiz, da für beide Übungen die Korrelationen zwischen dem jeweiligen Quizergebnis und der entsprechenden Leistungsprognose .54 bzw. .53 betragen. Die letzte Spalte listet die Validitätskoeffizienten der studentischen Leistungsprognose auf und lässt erkennen, dass diese sowohl auf der Basis der Quiz wie der Musterlösung im Mittel annähernd gleich ausfallen. D.h., auch mit Hilfe von Musterlösungen kann der Studierende sein Leistungsniveau ähnlich gut wie bei einer Testung einschätzen, obwohl es bei Musterlösungen weder eine Testung noch Rückmeldungen gibt. Diese These wird in Tabelle 13 weiter belegt.

Tabelle 13: Vergleich des Betrages der Abweichung von Schätzung und Lernerfolgsergebnis für Quiz und Musterlösungen und Wilcoxon Test für abhängige Stichproben

		Median
Übung 1:	ABS (JOL nach Quiz - Lernerfolg)	10.0
	ABS (JOL nach Muster - Lernerfolg)	11.8
	N=87; z= -0.88; p= 0.38	
Übung 2:	ABS (JOL nach Quiz - Lernerfolg)	14.5
	ABS (JOL nach Muster - Lernerfolg)	11.3
	N = 77; z = 1.94; p = 0.053	

Im Median verschätzten sich die Studierenden um 10 bis 15 Prozentpunkte nach oben oder unten. Zwischen den Übungsmethoden findet man kaum nennenswerte Unterschiede. Eine Inspektion der Mediane zeigt weiterhin auf, dass man zu ähnlichen Ergebnissen gelangen würde, wenn man die Daten auf der Basis unabhängiger Vergleiche geprüft hätte.

Insgesamt weisen die Ergebnisse eindeutig darauf hin, die Bearbeitung von Musterlösungen ziehe im Mittel keine höheren Kompetenzillusionen als die Bearbeitung der Quiz nach sich, sondern ermögliche eine mindestens so gute Einschätzung der eigenen Leistung wie die Bearbeitung von Tests mit Rückmeldungen. Da aber nur Tests das objektive Testergebnis rückmelden können - es hier nach der JOL-Schätzung auch taten -, kann der Studierende danach seine eigene Einschätzung validieren und gegebenenfalls ändern. Diese Studie sagt selbstverständlich nichts über einen Prognosevergleich von Musterlösungen gegenüber solchen Quiz aus, die eine JOL-Erhebung erst nach der Rückmeldung des Knowledge of performance durchführen.

Zusammenfassung und Diskussion

Vergleichbarer Lernerfolg für Testen mit Feedback und Musterlösungen

Alle Ergebnisse bestätigen die vergleichbare Lernwirkung von Musterlösungen und Testen mit Feedback, wobei Musterlösungen meist weniger Lernzeit als das Testen abverlangten und somit teilweise eine höhere Lerneffizienz aufwiesen. Ähnlich wie bei den früheren Studien von Jacobs (2008, 2010, 2011) hat sich die Musterlösung vom Standpunkt objektiven Lernerfolgs somit als echte Alternative zu Testen mit Feedback bewährt. Da eine reine Testgruppe ohne Feedback nicht verfügbar war, lässt der Untersuchungsansatz keine gesonderte Abschätzung des Quizerfolgs in die Effekte des Testens und die des Feedbacks zu. Aus den vergleichbaren Lernerfolgwerten von Quiz und Musterlösung darf jedoch gefolgert werden, ein Testen sei nicht notwendig gewesen und das Feedback sei in beiden Treatments in vergleichbarer Weise repräsentiert worden.

Auf den ersten Blick stehen die Ergebnisse im Widerspruch zu den Befunden vieler Laborstudien und einiger Feldstudien, weil sie den starken „testing effect“ - die Überlegenheit des Testens (mit Feedback) gegenüber der erneuten gezielten Präsentation der Information - nicht bestätigen (siehe dazu den Gesamtüberblick in der Metaanalyse von Rowland, 2014).

Andererseits unterscheidet sich das Untersuchungsszenario auch in etlichen Punkten von dem rein experimentellen Vorgehen bisheriger Studien (siehe dazu auch Anhang 1). Besonderer Wert wurde hier auf die Äquivalenz der Informationen von Quiz und Musterlösung gelegt. Alle Informationen in Musterlösung und Quiz sind vergleichbar. Der Unterschied liegt lediglich darin, beim Quiz die Aufgaben selbst zu beantworten, während in der Musterlösung die korrekten Antworten vorliegen. In manchen Studien fehlte bei der Musterlösung z.B. die Fragestellung, statt aller MC-Alternativen sahen die Probanden nur die korrekte Alternative, anstelle ganz gezielter Fragen sollte die Kontrollgruppe bestimmte Textpassagen erneut durchlesen.

Der Zeitabstand zwischen Instruktion und Übung fällt in Laborstudien meist sehr kurz aus, beträgt hier aber einige Tage. Im Gegensatz zu den meisten Laborstudien konnten die Probanden ihre Lernzeit selbst bestimmen. Ein intensives selbstinitiiertes Eigenstudium bzgl. der quizrelevanten Themen erscheint im Mittel zwar unwahrscheinlich, wurde aber nicht verunmöglicht.

Insgesamt war die Untersuchung in eine echte Vorlesung integriert und reflektiert somit einen pädagogisch realistischen Ausschnitt praktisch universitärer Lehre. Die Erfolgsquoten der Quiz von ca. 60% korrekter Lösungen fielen eher gering aus, um einen deutlichen Testeffekt zu produzieren, aber gering genug, um eine messbare Feedbackwirkung durch Korrektur der Fehler zu ermöglichen. Die das Behalten stabilisierende Wirkung des Testens zeigte sich vornehmlich dann, wenn eher Constructed Response Aufgaben verlangt und sowohl im Quiz wie im Behaltenstest identische Aufgaben verwendet wurden. Hier kamen bis auf eine Aufgabe nur Selected response Aufgaben, vornehmlich Zuordnungs- und True/False Aufgaben, zum Einsatz.

Die Aufgaben in Übung und Lernerfolgskontrolle waren nicht identisch, sondern ähnlich. Insofern forderten die Items des Lernerfolgstests gegenüber den Übungsaufgaben im Quiz bzw. der Musterlösung einen gewissen Transfer. Reine Memorierung reichte

meist nicht aus, um den Lernerfolg zu sichern und dann sind neben schwächeren Lernerfolgswerten auch geringere Treatmentunterschiede zu erwarten. Höchstwahrscheinlich schwindet der testing effect mit wachsendem Lehrzielanspruch. Möglicherweise kommt es unter den hier vorhandenen Bedingungen weniger auf das Übungsformat an, sondern auf die Bereitschaft, die in beiden Methoden präsentierten identischen Informationen hinreichend zu elaborieren. Wie die Worked-Example-Forschung aufzeigte, profitieren Novizen eher von Lösungsbeispielen und erst Experten von eigenen Lösungsversuchen bzw. einer Testung. In Mittelbereich eines noch nicht hinreichend konsolidierten Wissens sind vermutlich nur geringe Unterschiede zwischen Testen mit Feedback und Musterlösungen zu erwarten.

Höhere Bewertung und Präferenz für Testen mit Feedback

Trotz vergleichbaren Lernerfolgs von Quiz und Musterlösung sowie einer höheren Lerneffizienz der Musterlösung bewerteten Studierende das Testen mit Feedback hinsichtlich seiner Lernwirksamkeit und Lernanregung eindeutig besser als die Musterlösung. Sie würden bei freier Methodenwahl das Testen mit Feedback der Bearbeitung von Musterlösungen vorziehen, auch wenn das Testen eine etwas höhere Beanspruchung erforderte. Möglicherweise geht ein Großteil der höheren Wertschätzung der Quiz gegenüber den Musterlösungen auf die objektive Rückmeldung des Gesamtergebnisses - Feedback of performance - zurück. Sie gewährt dem Studierenden eine verlässliche Auskunft über seinen Leistungsstand. Ohne „Feedback of performance" lieferten beide Übungsmethoden aber eine ähnlich gute Grundlage, die eigene Leistung einzuschätzen. D.h., auch auf der Basis einer Musterlösung konnte der Studierende eine ähnlich gute Prognose seiner Leistung in einer Woche abgeben wie mit Hilfe der Testung mit lediglich Aufgaben spezifischem Feedback. Vermutlich gelingt es dem Studierenden bei Sichtung der Musterlösung ungefähr abzuschätzen, wie gut er eine analoge Testfrage selbst hätte beantworten können.

Übungen besser als keine Übung

Ob die Übungen besser als keine Übung sind, konnte in den bisherigen Studien von Jacobs (2008, 2010, 2011) gar nicht geklärt werden, da stets eine adäquate Kontrollgruppe fehlte. Der Autor mahnte zu Recht an, man könne aus der vergleichbaren Lernwirkung von Quiz und Musterlösung keineswegs auf deren prinzipielle Lernwirksamkeit schließen. In vorliegender Studie standen zwei Kontrollgruppen zur Verfügung, aber nur der Vergleich der Übungsgruppen mit der durch die Randomisierung kontrollierten Kontrollgruppe liefert ein hinreichend vertrauenswürdiges Ergebnis. Danach erbrachten sowohl die Übung mit Hilfe von Musterlösungen wie auch das Quiz mit Feedback einen statistisch gesicherten Lerneffekt gegenüber gar keiner Übung, wenn auch im niedrigen Effektstärkebereich. Jacobs (2009, 2010b) fand ähnlich mäßige Lernerfolge des Testens mit Feedback gegenüber gar keiner Übungsmethode. Häufig suggerieren unkontrollierte Kontrollgruppen höhere Effekte (siehe hier der Vergleich mit der KG' oder die riesige Effektstärke bei Jacobs, 2008), aber nur auf Kosten mangelnder interner Validität. Der vom Ausmaß her eher bescheidene, aber wegen der strengen Prüfung und der hohen pädagogischen Anforderungen dennoch bemerkenswerte Lernvorteil beider Übungsmethoden fällt erkennbar schwächer aus als in etlichen Laborstudien.

Dafür könnten mehrere Faktoren verantwortlich sein. Zum einen hatten alle Studierenden, also auch die Kontrollgruppe, jederzeit Zugriff auf die lehrzielrelevanten Informa-

tionen und so die Möglichkeit, den vorausgegangen Lehrstoff nachzuarbeiten. Es gab jedoch wenig ersichtlichen Grund dies zu tun, weil die Studierenden den Lernerfolgstest gar nicht erwarteten, was möglicherweise ihre Übungsmotivation dämpfte. Außerdem wurden in der Vorlesung die Themen der jeweils vorherigen Vorlesungssitzung in ihren wichtigsten Passagen erneut dargeboten. Etliche Studierende der Kontrollgruppe konnten so zumindest für Lernerfolg 1 von einer Stoffwiederholung profitieren und den Übungsvorteil der Treatmentgruppen teilweise vermindern.

Feedbackzeitpunkt spielt keine Rolle

Fast alle Messvariablen ergaben hoch vergleichbare Testwerte für beide Feedbacktimingvarianten. Ob die Rückmeldungen unmittelbar nach jeder Aufgabe oder erst nach Bearbeitung aller Aufgaben einsetzten, spielte somit überhaupt keine Rolle im Hinblick auf die Quizleistung, die Quizbearbeitungszeit, die Quizbewertungen und den Lernerfolgstest. Die Studie bestätigte hinsichtlich des Lernerfolgstests somit die Ergebnisse von Buzhardt & Semp (2002), Henshaw (2011) und van der Kleij et al. (2012). Da nur unter Q_EOTF die Bearbeitungszeit der Testung von der des Feedbacks getrennt erfasst werden konnte, ist ein Feedbacktimingvergleich mit der Studie von van der Kleij (2012) nur unter der Annahme möglich, bei gleichen Gesamtbearbeitungszeiten beider Quiz hätten sich die Zeiten für das Testen und das Feedback auch in ähnlicher Weise aufgeteilt. Das würde dann den Befunden von van der Kleij et al. (2012) widersprechen, die von signifikant höheren Feedbackbearbeitungsarbeitszeiten für unmittelbares Feedback berichten. Des Weiteren hätten Studierende dem unmittelbaren Feedback eigenen Angaben zu Folge mehr Aufmerksamkeit geschenkt, was sich aber nicht im Lernerfolg widerspiegelte. Die hier eingesetzten Quiz waren mit 7 bis 8 etwas umfangreicheren Aufgaben relativ kurz, da die Quizbearbeitung einschließlich des Feedbackstudiums im Median nur ca. 14 Minuten in Anspruch nahm. Möglicherweise sinkt die Feedbackbearbeitungszeit des EOTF im Vergleich zum FAI mit wachsender Testlänge, weil das Feedback dann erst am Ende erscheint, wenn der Zeitdruck besonders groß ist.

Einige Bemerkungen zur Generalisierung der Befunde

Eine einfache Übertragung der Ergebnisse auf die vielfältigen Konstellationen der Schulwirklichkeit erweist sich als schwierig und muss die besonderen Bedingungen dieser Studie berücksichtigen. Sie besagen im Wesentlichen: Wenn sich Studierende nach einer mehr oder weniger umfangreichen Instruktionsphase in halbwegs kontrollierter Weise dazu verpflichtet sehen, eine Übung einigermaßen konzentriert zu bearbeiten, dann können sie von Quiz mit Feedback in ähnlicher Weise profitieren wie von Musterlösungen. Der Übungsgewinn beider Treatmentvarianten sollte ähnlich ausfallen, aber Musterlösungen gewisse Zeitvorteile aufweisen.

Aber Studierende werden z.B. in der Regel selten, also anders als hier im Unklaren darüber gelassen, welche Art von Übung sie erwartet. Wissen Sie im Vorhinein, dass Musterlösungen anstehen, so könnten sie sich im Vergleich zu einem Test vermutlich weniger gut darauf vorbereiten. Bei Onlineübungen besteht häufig - im Gegensatz zu hier - die Möglichkeit, Übungen beliebig oft zu wiederholen, was mit zunehmender Erfolgsquote eher eine Testung begünstigt. Etliche Arbeitsaufträge stehen zur Bewertung an, aber Musterlösungen lassen sich nicht bewerten, es sei denn, man kündigt einen Test zur Überprüfung der Musterlösungsbearbeitung an. Bei freiwilligen, unbenoteten Übungen während des Semesters werden Studierende vermutlich eher Quiz

mit Feedback anwählen, obwohl Musterlösungen häufig günstiger wären. Musterlösungen erlauben zwar auch eine mehr oder weniger gute Leistungsprognose, geben aber letztendlich keine objektive Rückmeldung zum aktuellen Leistungsstand. Unter Zeitdruck besteht die Gefahr, sich für Musterlösungen weniger Zeit zu nehmen (siehe Tabelle 4), aber für eine vom Studierenden selbst initiierte, gezielte Prüfungsvorbereitung würde dieser vermutlich gerne auf lehrzielrelevante Musterlösungen zugreifen, weil diese in kürzerer Zeit als Tests zu bearbeiten sind. Da Musterlösungen sich vollständig aus den Informationen der Aufgaben und des Feedbacks zusammensetzen, wäre es ohne großen technischen Aufwand möglich, beide Varianten standardmäßig anzubieten.

Literatur

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples:Instructional principles from the worked examples research. Review of Educational Research, 70, 181-214.

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

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.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1118–1133.

Brosvic, M. & Epstein, M. L. (2007). Enhancing learning in the introductory course. Psychological Record, 57, 391-408.

Brosvic,G.M., Epstein, M.L. Cook, M.J., Dihoff.. R.E. (2005). Efficacy of Error for the Correction of Initially Incorrect Assumptions and of Feedback for the Affirmation of Correct Responding: Learning in the Classroom. Psychological Record, 55, 401-418.

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

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. Applied Cognitive Psychology, 14, 215-235.

Buzhardt, J & Semp, G.B. (2002). Item-by-Item Versus End-of-Test Feedback in a Computer-Based PSI Course Journal of Behavioral Education, 11, [2] 89–104.

Dihoff, R. E.; Brosvic, G. M., Epstein, M. L. (2003) The Role of Feedback During Academic Testing: The Delay Retention Effect Revisited. *The Psychological Record*. 53, 533-548.

Dihoff, R. E.; Brosvic, G. M., Epstein, M. L., Cook, M.J. (2004). Provision of Feedback During Preparation for Academic Testing: Learning is Enhanced by Immediate but not Delayed Feedback. The Psychological Record, 54, 207-231.

Henshaw, F. G. (2011). Effects of Feedback Timing in SLA: A Computer-Assisted Study on the Spanish Subjunctive. in. Sanz, C. & Leow R. P., 2011. Implicit and explicit language learning. Conditions, Processes, and Knowledge in SLA.Georgetown University Press.

Howard, Ch, R (2010). Examining the Testing Effect in an Introductory Psychology Course
A dissertation submitted to the Graduate Faculty of Auburn University, Auburn, Alabama
http://etd.auburn.edu/etd/bitstream/handle/10415/2289/Dissertation%20Manuscript%20%28Final%20Draft%29.pdf?sequence=2

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. (2008).Gezieltes Studieren gelöster Aufgaben als alternative
Übungsmethode zu Testen mit Feedback..
URN: urn:nbn:de:bsz:291-psydok-15597
URL: http://psydok.sulb.uni-saarland.de/volltexte/2008/1559/

Jacobs, B. (2009). Leistungssteigerung durch Notendruck? - Die Wirkung der Benotung
auf die Studierleistungen in einem Seminar.
URN: urn:nbn:de:bsz:291-psydok-25299
URL: http://psydok.sulb.uni-saarland.de/volltexte/2009/2529/

Jacobs, B. (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. (2010b). Leistungssteigerung ohne Notendruck ? -Die Wirkung verpflichtender,
unbenoteter Quiz auf die Studierleistung.
URN: urn:nbn:de:bsz:291-psydok-26899
URL: http://psydok.sulb.uni-saarland.de/volltexte/2010/2689/

Jacobs, B. (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/

Kang, S.K., McDermott, K.B., & Roediger, H.L. (2007). Test format and
corrective feedback modulate the effect of testing on memory retention.
European Journal of Cognitive Psychology, 19, 528-558.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal
learning. Review of Educational Research, 58, 79–97.

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

Larsen, D. P., Butler, A. C., Roediger III, H. L. (2009). Repeated testing improves
long-term retention relative to repeated study: a randomised controlled trial. Medical Education, 43,
1174–1181.

Paas, F. G. W. C., & Van Merrienboer, J. J. G. (1994). Variability of worked examples and
transfer of geometrical problem-solving skills: A cognitive-load approach.
Journal of Educational Psychology, 86(1), 122-133.

Pilotti, M. , Chodorow, M. & Petrov, R. (2009): The Usefulness of
Retrieval Practice and Review-Only Practice for Answering Conceptually Related Test Questions,
The Journal of General Psychology, 136:2, 179-204

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.

McDaniel, M.A.; Wildman, K.M. & Anderson, J. L. (2012).
Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. Journal of Applied Research in Memory and Cognition. 1 [1] 18-26.

Roediger, H. L. ,& Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science, 1, 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17, 249–255.

Roediger, H.L., Agarwal, P.K., McDaniel, M.A., McDermott, K.B. (2011).
Test-enhanced learning in the classroom: Long-term improvements from quizzing.
Journal of Experimental Psychology: Applied. 17 [4], 382–395

Rowland, C. A. (2014). The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect. Psychological Bulletin. Advance online publication.
http://dx.doi.org/10.1037/a0037559-

Toppino T. C, Cohen M. S. (2009). The testing effect and the retention interval: questions and answers. Experimental Psychology, 56 (4), 252-257.

van der Kleij, F.M. , Eggen, T.J.H.M., Timmers, C.F. & Bernard P. Veldkamp, B.P. (2012).
Effects of feedback in a computer-based assessment for learning. Computers & Education 58, 263–272

Anhang 1:

Überblick wichtiger Bedingungen für Übung und Lernerfolgsmessung	
Teilnahme am Experiment	freiwillig, unterstützt durch marginalen leistungsunabhängigen Bonus für die Abschlussklausur
Übungsgrundlage /Instruktion	vorausgehende 2 Vorlesungen und Powerpoint.-Folien des Dozenten.
Übungsankündigung	nicht als solche bezeichnet, sondern als Erhebung. Studierende konnten nicht antizipieren, dass eine Übung anstand.
Übungsrelevante Lehrziele	den Studierenden zuvor nicht expliziert: überwiegend Verständnis und Anwendung
Übungsmethode	für Studierende nicht vorhersehbar: Quiz mit Feedback sowie Musterlösung.
Übungsdurchführung online zu Hause	mehrere Tage nach der letzten Vorlesungssitzung; bei Quiz und Musterlösung musste jede Aufgabe bestätigt werden. Self-paced: Es gab weder Mindestzeiten noch Zeitbegrenzungen.
Lernerfolgsmessung streng kontrolliert im Vorlesungssaal.	7 bis 12 Tage nach der Übung; für Studierende unerwartet parallele Aufgaben zur jeweiligen Übung
Übungs- und Lernerfolgsbewertung	Stets unbenotet, da anonym Übung: mit milder Zufallskorrektur Lernerfolg: ohne Zufallskorrektur