Unterrichtswiss (2022) 50:37–52
https://doi.org/10.1007/s42010-021-00137-4

THEMENTEIL

Komplexität und der Testungseffekt: Die mögliche Bedeutung der Verständnissicherung für den Nutzen von Abrufübung bei komplexem Lernmaterial

Ralf Rummer · Judith Schweppe

Eingegangen: 20. April 2021 / Überarbeitet: 31. Juli 2021 / Angenommen: 19. Oktober 2021 / Online publiziert: 10. November 2021

Zusammenfassung Testung im Sinne eines aktiven Abrufs von Informationen aus dem Langzeitgedächtnis gilt als eine der effektivsten Möglichkeiten, Wissen zu konsolidieren und so nachhaltiges Lernen zu befördern. Der Testungseffekt gilt als robust und wurde für unterschiedlichste Personengruppen und Lernmaterialien gezeigt. Allerdings wird immer wieder kontrovers diskutiert, inwieweit der Testungseffekt auch bei komplexen Lernmaterialien auftritt. Der vorliegende Beitrag reflektiert diese Debatte. Dabei wird zunächst die theoretische Position derer nachvollzogen, die den Testungseffekt vor allem auf wenig komplexe Materialien beschränkt sehen. Diese Position wird anschließend anhand einer Problematisierung des Komplexitätsbegriffs und seiner Operationalisierung kritisch diskutiert. Schließlich wird eine alternative Erklärung für das potenzielle Fehlen des Testungseffekts bei komplexen Materialien skizziert, nach der das Auftreten des Testungseffekts nur indirekt von der Komplexität des Lernstoffs bzw. Lernmaterials abhängt. Gemäß dieser Annahme ist die Voraussetzung für das Auftreten des Testungseffekts, dass der Lernstoff während des initialen Lernens (also der Phase, die der Testung vorausgeht) hinreichend gut verstanden wurde und entsprechend Informationen im Langzeitgedächtnis enkodiert wurden, deren Abruf dann in einer Testungsphase geübt werden kann. Dies kann bei komplexen Materialien eine längere initiale Lernphase oder andere Maßnahmen der Verständnissicherung erfordern als bei einfachen Materialien. Abschließend wird skizziert, wie diese Annahme experimentell überprüft werden kann und welche prak-

Die Autoren R. Rummer und J. Schweppe teilen sich die Erstautorenschaft.

Ralf Rummer (✉)
Allgemeine Psychologie, Institut für Psychologie, Universität Kassel, Holländische Str. 36–38, 34127 Kassel, Deutschland
E-Mail: rummer@uni-kassel.de

Judith Schweppe
Psychologie mit Schwerpunkt Lehren und Lernen mit digitalen Medien, Universität Passau, Innstraße 41, 94032 Passau, Deutschland
E-Mail: judith.schweppe@uni-passau.de

tischen Implikationen sich daraus für eine möglichst lernwirksame Umsetzung von Abrufübung selbst mit komplexen Lernmaterialien ergeben.

Schlüsselwörter Testungseffekt · Komplexität von Lernmaterialien · Open-Book-Tests · Cognitive Load Theory · Elementinteraktivität

Complexity and the testing effect: The potential importance of securing understanding for the benefits of retrieval practice with complex learning materials

Abstract Testing (i.e., active retrieval of information already encoded in long-term memory) is considered as one of the most effective ways to promote lasting learning. The testing effect is robust and applies to various groups of learners and learning materials. However, since the testing effect has been discovered in the early 20th century, it has been discussed whether it also applies to highly complex learning materials. The present article reiterates and reflects this debate. In particular, it discusses the conceptualization and operationalization of complexity, and the empirical evidence for the assumption of complexity as a boundary condition of the testing effect. An alternative explanation for the potential lack of a testing effect with complex learning matters is outlined, according to which the occurrence of the testing effect is conceptualized as largely independent of the complexity of the learning material. According to this conception, the prerequisite for the occurrence of the testing effect is that the learning material has been understood sufficiently well during initial learning (i.e., the phase preceding testing) so that learners have successfully encoded the information in long-term memory, the retrieval of which they are supposed to practice in a testing phase. For complex materials, this may require a longer initial learning phase than for simple materials. Finally, we make suggestions as to how this assumption can be tested and outline practical implications for an efficient implementation of retrieval practice even with complex learning materials.

Keywords Testing Effect · Complexity of Learning Material · Open-Book-Tests · Cognitive Load Theory · Element Interactivity

1 Einleitung

Bis vor kurzem gab es in den einschlägigen deutschsprachigen Wörterbüchern der Psychologie (z. B. dem Dorsch; Wirtz 2013) *keinen* Eintrag unter dem Stichwort „Testungseffekt“ oder „Testeffekt“, obwohl dieser einerseits seit langem bekannt ist (z. B. Abott 1909; Gates 1917; Kühn 1914; Witasek 1907) und das Testen andererseits als eine der wirksamsten Methoden gilt, um nachhaltiges, also langfristiges Lernen zu verbessern (Roediger und Karpicke 2006a, b; Rummer et al. 2017; für umfassende Übersichten oder Metaanalysen siehe z. B. Pan und Rickard 2018; Rowland 2014; Rummer 2021; Schwieren et al. 2017; Yang et al. 2021). Konkret besagt dieser Testungseffekt (*Testing Effect*), der im englischen Sprachraum auch *Quizzing Effect*

oder *Retrieval Practice Effect* heißt, dass der aktive Abruf von Lerninhalten aus dem Langzeitgedächtnis die abgerufene Information konsolidiert und damit längerfristiger verfügbar hält, als dies bei Nutzung anderer Vorgehensweisen wie wiederholtem passiven Lernen der Fall wäre. Konkret heißt dies, dass der Testungseffekt sich vor allem langfristig zeigt, also dann, wenn der zeitliche Abstand zwischen der Lern- bzw. Konsolidierungsphase und dem kritischen finalen Test zumindest einige Tage dauert. Bisherige Daten sprechen dafür, dass der Effekt umso größer ist, je länger dieses Behaltensintervall ausfällt (Rowland 2014).

Experimente zum Testungseffekt sehen üblicherweise so aus, dass den Lernenden zunächst Lernmaterialien (z. B. Wortpaare oder expositorische Texte) vorgelegt werden. Diese initiale Lernphase ist in der Regel für alle Experimentalgruppen (also die Testungsgruppe(n) und Kontrollgruppe(n)) gleich. Die experimentelle Manipulation besteht nun darin, dass man die Versuchspersonen in der Testungsgruppe instruiert, das Lernmaterial wiederzugeben, bzw. offene oder geschlossene Fragen zum Material zu beantworten. In der Kontrollgruppe werden die Versuchspersonen hingegen standardmäßig instruiert, den Lernstoff erneut passiv zu wiederholen. Bisweilen werden in diesen Experimenten aber auch andere aktive Lernmethoden wie das Erstellen von *Concept-Maps* (Blunt und Karpicke 2014; Karpicke und Blunt 2011) oder Notizenmachen (McDaniel et al. 2009; Nguyen und McDaniel 2016; Rummer et al. 2017) als Kontrollbedingungen in die Untersuchungen integriert. In der Regel zeigen diese Experimente kurzfristig – typischerweise nach einem Behaltensintervall von ca. 5 min – ein uneinheitliches Bild. Meist findet sich kein Testungseffekt und mitunter sogar ein Nachteil für Testen. Langfristig ändert sich die Befundlage, und es findet sich nach wenigen Tagen ein deutlicher Vorteil des Testens gegenüber passivem Wiederholen und nach zwei Wochen auch gegenüber aktiven Kontrollbedingungen wie Notizenmachen (Rummer et al. 2017, Exp. 1).

Auch in Feldexperimenten in Schule und Hochschule hat sich Testung als wirksam erwiesen. Hier werden typischerweise über einen längeren Zeitraum Übungstests zu verschiedenen Lektionen bearbeitet. In einer exemplarischen Studie variierten Batsell et al. (2016) zwischen zwei Einführungskursen in die Psychologie, ob die Studierenden zu Leseaufgaben begleitende Übungstests bekamen oder nicht. In der Kursprüfung schnitt die Testungsgruppe besser ab als die Kontrollgruppe, und zwar sowohl wenn die Fragen die gleichen waren wie die bereits geübten als auch bei neuen Fragen. In Feldexperimenten finden sich häufiger derartige Kontrollbedingungen ohne vergleichbare Lernaktivität. Eine aktuelle Metaanalyse zeigt aber auch gegenüber anderen Kontrollbedingungen – wie dem in Laborexperimenten typischerweise eingesetzten wiederholten Lesen – einen substantiellen Testungseffekt im Feld (Yang et al. 2021). Aufgrund der Vielzahl von Befunden, die für eine lernförderliche Wirkung sprechen, wurde Testen in mehreren Überblicksarbeiten zu effektiven Lern- und Lehrstrategien aufgenommen (z. B. Dunlosky et al. 2013; Pashler et al. 2007) und als Methode herausgestellt, deren Wirksamkeit besonders umfassend und in unterschiedlichsten Kontexten geprüft wurde. Dies reflektieren auch die zahlreich vorliegenden Metaanalysen zum Testungseffekt, die übereinstimmend mindestens mittlere Effektstärken nachwiesen, sowohl in Bezug auf Laborexperimente (Rowland 2014; Pan und Rickard 2018) als auch hinsichtlich Studien in authentischen Lernsituationen in Schule und Hochschule (Schwieren et al. 2017; Yang et al. 2021).

2 Direkte und indirekte Testungseffekte

Die Nützlichkeit von Übungstests wird sowohl auf direkte als auch auf indirekte Mechanismen zurückgeführt. Basierend auf dieser Unterscheidung wird auch begrifflich zwischen einem direkten und einem indirekten Testungseffekt unterschieden, wobei der direkte Testungseffekt sich aus einer konsolidierenden Wirkung des Tests auf die entsprechenden Repräsentationen ergibt, während der indirekte Testungseffekt sich auf nachfolgende Enkodierprozesse bezieht.

2.1 Der direkte Testungseffekt

Als Erklärung für einen direkten Testungseffekt wird angenommen, dass der Akt des Abrufs von Informationen aus dem Langzeitgedächtnis den späteren Abruf dieser Informationen verbessert, indem die Gedächtnisspur gestärkt wird. Zudem wird angenommen, dass durch den aktiven und aufwändigen Gedächtnisabruf während der Übungstests neue Abrufrouten etabliert und bestehende gestärkt werden. Hierdurch wird dann der langfristige Zugang zu diesen Informationen verbessert (z. B. Bjork 1975; Bjork und Bjork 2011; Rowland und DeLosh 2014). Carpenter (2009) geht darüber hinaus davon aus, dass beim Wissensabruf auch Repräsentationen aktiviert werden, die semantisch mit dieser Information assoziiert sind. Demzufolge profitieren auch Informationen, die nicht direkt abgerufen wurden (*Elaborative Retrieval Hypothese*, s. auch Carpenter und DeLosh 2006). Entsprechend dieser Überlegungen ist die Art des Abrufs essentiell für den Nutzen der Testung. Je höhere Anforderungen mit dem Wissensabruf einhergehen (z. B. Wiedergabe vs. Rekognition), desto effektiver und nachhaltiger wirkt Testung. Ein weiterer Erklärungsansatz für den direkten Testungseffekt, der die Bedeutung des Abrufaktes betont, stützt sich auf die sog. transferangemessene Verarbeitung (transfer-appropriate processing, Morris et al. 1977; Roediger und Karpicke 2006a). Die basale Annahme ist, dass Erinnerungs- und Transferleistungen umso besser gelingen, je mehr die während der Enkodierung ausgeführten Prozesse denen während des Abrufs ähneln. Bezogen auf den Testungseffekt bedeutet dies also, dass das Üben des Abrufs während des Lernens deshalb vorteilhaft ist, weil es bereits früh im Lernprozess die Prozesse beansprucht, die dann, wenn man sich zu einem späteren Zeitpunkt an diese Information erinnern will, ebenfalls benötigt werden (einen ausführlichen Überblick über die Grundlagenforschung zur Erklärung des Testungseffekts geben Tempel und Pastötter 2021).

2.2 Der indirekte Testungseffekt

Neben diesen direkten Aspekten kann Testung das Lernen auch auf indirekte Weise verbessern, und zwar indem nachfolgende Enkodierprozesse verbessert bzw. zielgenauer ausgerichtet werden (z. B. Endres und Renkl in diesem Themenheft; Roediger und Karpicke 2006a; Rummer et al. 2019; Rummer 2021; Tempel und Pastötter 2021). Hier wird auch von test-potenziertem Lernen gesprochen (Arnold und Mc-

Dermott 2013; Tempel und Frings 2019).[1] Im Gegensatz zum wiederholten Lesen, bei dem Lernende leicht den Eindruck haben können, den Lernstoff aufgrund des sich mit jedem Lesen verstärkenden Vertrautheitseindrucks sehr gut zu beherrschen, auch wenn dies de facto nicht der Fall ist, können Übungstests Wissenslücken oder nicht leicht zugängliches Wissen aufdecken und somit metakognitiv fassbar machen. Infolgedessen sollten Lernende, die getestet werden, sowohl besser entscheiden können, ob und wie lange sie sich erneut mit dem Lernstoff befassen, als auch, welche Informationen sie bei der erneuten Beschäftigung mit dem Material fokussieren. Die positive Wirkung von Testung entsteht somit nicht (nur) während des Testens selbst, sondern während der folgenden erneuten Auseinandersetzung mit dem Lernmaterial oder mit Teilen davon. Eine Voraussetzung für den indirekten Testungseffekt ist daher naturgemäß, dass die Lernenden im Anschluss an einen Übungstest in irgendeiner Form die Möglichkeit haben, sich erneut mit dem Lernstoff zu befassen. Zudem sollten diese Vorteile vor allem in Lernsituationen zum Tragen kommen, in denen die Lernenden motiviert sind, entdeckte Wissenslücken tatsächlich zu schließen oder nicht abrufbares Wissen durch die erneute Auseinandersetzung mit dem Lernmaterial besser verfügbar zu machen. Entsprechend argumentieren Yang et al. (2021) im Rahmen ihrer Metaanalyse, dass dem indirekten Testungseffekt vor allem im Feld eine bedeutende Rolle zukommt.

3 Randbedingungen für das Auftreten des Testungseffekts

Testungseffekte treten weitgehend unabhängig von interindividuellen Unterschiedsvariablen wie Alter oder Arbeitsgedächtniskapazität auf (z. B. Bertilsson et al. 2021; für einen Überblick s. Dunlosky et al. 2013). Außerdem zeigen sich Testungseffekte mit einer Reihe unterschiedlicher Materialien und in Feldexperimenten unabhängig vom Fach (Yang et al. 2021). Sowohl die bisherigen Metanalysen (z. B. Rowland 2014; Yang et al. 2021) als auch weitere Einzelstudien weisen allerdings darauf hin, dass die Ausprägung des Testungseffekts von einigen Randbedingungen abhängt. Wie bereits oben skizziert, ist Testung beispielsweise effektiver, wenn der finale Test nicht unmittelbar auf die Lernphase folgt, sondern zeitlich um einige Tage versetzt stattfindet. Außerdem ist Testung effektiver, wenn die Lernenden das Lernmaterial nach der Abrufübung wiederholen können oder in anderer Form Feedback erhalten, als wenn dies nicht der Fall ist. Dies zeigt sich sowohl im Labor (Rowland 2014) als auch im Feld (Yang et al. 2021) und kann unter anderem damit erklärt werden, dass in diesem Fall sowohl der direkte als auch der indirekte Testungseffekt zum Tragen kommen können. Schließlich sind nicht alle Testverfahren gleichgut geeignet, um die Lernleistung zu konsolidieren. So sind beispielsweise Wiedererkennenstests oder Multiple-Choice-Aufgaben weniger effektiv als freie Wiedergabe oder die Beantwortung offener Fragen (Rowland 2014). Im Feld sieht die Situation hier allerdings möglicherweise anders aus – so zeigte die Metaanalyse von Yang

[1] Hiervon wiederum abzugrenzen ist das test-potenzierte neue Lernen oder der „forward testing effect", unter dem positive Effekte von Tests auf das Lernen folgender, neuer Informationen verstanden werden (s. z. B. Chan et al. 2018; Pastötter und Bäuml 2014; Yang et al. 2018).

et al. (2021) vergleichbar große Testungseffekte für Wiedererkennens- und Abrufaufgaben (s. a. Greving und Richter 2018).

Es gibt ebenfalls Studien, die darauf hinweisen, dass der Testungseffekt mit der Anzahl der durchgeführten Tests größer wird (für einen Überblick s. Dunlosky et al. 2013). So verglichen Roediger und Karpicke (2006b, Exp. 2) eine Bedingung, in der Lernende einen kurzen Text in vier aufeinanderfolgenden Phasen wiederholt lesen konnten, eine Bedingung, in der auf drei Lesephasen eine Testphase folgte, und eine Bedingung, in der auf eine einzige Lesephase drei Testphasen folgten. Während die Leistungen im unmittelbaren Abschlusstest umso besser waren, je mehr Lesephasen vorgesehen waren, kehrte sich das Muster um, wenn der Abschlusstest eine Woche später durchgeführt wurde. Hier übertrafen die Lernenden mit mindestens einer Testphase diejenigen, die ausschließlich durch Lesen gelernt hatten, aber auch die Lernenden mit nur einer Lesephase und drei Testphasen übertrafen die Lernenden mit drei Lesephasen und einer einzigen Testphase. Auf einen weiteren möglichen Moderator, die Komplexität des Lernmaterials, wollen wir im Folgenden detaillierter eingehen.

4 Die Komplexität des Lernmaterials: ein Moderator des Testungseffekts?

Dafür, dass der (direkte) Testungseffekt hinsichtlich seiner Ausprägung davon abhängt, wie komplex das Lernmaterial ist, argumentierten van Gog und Sweller (2015). Mehr noch: Die Autor*innen stellten sogar fest, dass der Testungseffekt auf wenig komplexe Lernmaterialien beschränkt ist. Das entsprechende Argument entwickelten sie in einem Sonderheft der Zeitschrift *Educational Psychology Review.* Auf Basis früher experimenteller Arbeiten (z. B. Kühn 1914; Gates 1917) und in dem Sonderheft versammelter aktueller Studien folgerten sie, dass der Testungseffekt auf Lernmaterialien geringer Komplexität begrenzt ist. Komplexität definierten sie dabei als die Anzahl der Elemente und deren Verbindungen untereinander, die Lernende gleichzeitig im Arbeitsgedächtnis halten müssen, um die für das Verständnis des Lerninhalts notwendigen Inferenzen zu ziehen. In der *Cognitive Load Theory* wird dies als Elementinteraktivität bezeichnet (z. B. Paas et al. 2003). Materialien mit einer geringen Elementinteraktivität sind beispielsweise Wortlisten oder Vokabeln (s. bereits Gates 1917). Jedes einzelne Item oder Vokabelpaar steht hier für sich und muss weder auf eine andere Information bezogen noch müssen Inferenzen gezogen werden. Materialien mit hoher Elementinteraktivität sind hingegen Materialien, deren Verständnis von den Relationen unterschiedlicher Aussagen zueinander abhängt. Ein Beispiel hierfür wäre die Rückbeziehung von Experimentalergebnissen auf die Gültigkeit einer Theorie. Hier muss ein Abgleich zwischen Theorie und Hypothesen, Ergebnissen und Hypothesen und Hypothesen und Theorie erfolgen.

Van Gog und Sweller (2015) kategorisierten eine Reihe von einschlägigen Experimenten zum Testungseffekt bezüglich der Komplexität der genutzten Lernmaterialien. Der Eindruck, den diese Übersicht vermittelt, ist eindeutig: Nur in zwei der 15 Studien mit komplexen Materialien zeigte sich ein Testungseffekt, verglichen mit 29 von 33 Studien mit einfacheren Materialien. Dies legt nahe, dass das Auf-

treten des Testungseffekts tatsächlich auf wenig komplexen Lernstoff mit niedriger Elementinteraktivität begrenzt ist.

Warum sollte dies der Fall sein? Van Gog und Sweller (2015) nehmen Bezug auf Erklärungen, die den direkten Testungseffekt darauf zurückführen, dass durch den aktiven Informationsabruf Beziehungen zwischen Elementen in stärkerem Maß etabliert oder verstärkt werden als bei wiederholtem Lesen. Laut ihrer Argumentation sollte dies vor allem bei solchen Materialien nützlich sein, die es nicht ohnehin erfordern, mehrere Elemente miteinander zu verbinden, also bei Materialien mit niedriger Elementinteraktivität. Komplexe Materialien mit hoher Elementinteraktivität erfordern es dagegen, auch ohne Abrufinstruktion, verschiedene Elemente miteinander in Beziehung zu setzen, sodass hier kein zusätzlicher Gewinn durch Testung zu erwarten wäre.

5 Was macht die Komplexität von Lernmaterialien aus?

Bereits im *Special Issue* äußern die eingeladenen Diskutanten Kritik am Vorgehen von van Gog und Sweller (2015; s. Karpicke und Aue 2015; Rawson 2015). Karpicke und Aue (2015) kritisierten die Subjektivität der Post-hoc-Kategorisierung der Lernmaterialien in den bereits publizierten Studien in hohe, mittlere oder niedrige Elementinteraktivität. Tatsächlich bleiben die Klassifikationskriterien weitgehend unklar, und bei vielen Studien haben van Gog und Sweller (2015) ihre Einstufung mit Fragezeichen versehen. Karpicke und Aue (2015) schlugen deshalb eine Kategorisierung auf Basis von etablierten Textkomplexitäts- und Textkohäsionsmetriken wie *Flesch reading ease* (Flesch 1948) oder *Coh-Metrix* (Graesser et al. 2014) vor. Der Flesch Lesbarkeitsindex ergibt sich aus der durchschnittlichen Satz- und Wortlänge; anhand von *Coh-Metrix* lässt sich (im Englischen) etwa die referentielle Kohäsion bestimmen, das Ausmaß, in dem Ideen in einem Text und entsprechend zwischen Sätzen überlappen, was laut Karpicke und Aue (2015) ein besonders gutes objektives Maß für Elementinteraktivität sein könnte. Die Analyse von Texten auf Basis dieser Metriken führte zu anderen Komplexitätszuschreibungen als den von van Gog und Sweller (2015) vorgeschlagenen. Diese Inkonsistenz macht deutlich, dass zunächst näher bestimmt werden sollte, was komplexe Lernmaterialien von einfachen unterscheidet, bevor Aussagen zur Auswirkung von Komplexität auf den Nutzen von Abrufübung getroffen werden können. Ein Problem der von Karpicke und Aue (2015) herangezogenen Metriken besteht allerdings darin, dass sie sich ausschließlich auf Texte beziehen und damit ein Großteil der im Sonderheft versammelten Studien gar nicht klassifiziert werden kann, da sich diese mit mathematischem Problemlösen befassen. Das zentrale Problem des Vorgehens von Karpicke und Aue (2015) ist aus unserer Sicht hingegen ein anderes und liegt in dem Umstand, dass durch die Gleichsetzung von Komplexität und *Text*komplexität die Komplexität der vermittelten Lern*inhalte* gar nicht berücksichtigt wird.

Dies führt zu einem Kernproblem der Komplexitätserfassung und -manipulation, nämlich zur Frage, ob Elementinteraktivität eine Eigenschaft des Lernstoffs oder eine Eigenschaft der instruktionalen Umsetzung, etwa der Formulierung des Lerntexts darstellt bzw. in der Terminologie der *Cognitive Load Theory* (z. B. Paas et al. 2003),

ob sich Elementinteraktivität auf intrinsische oder auf extrinsische kognitive Belastung bezieht. Diese Frage mag zunächst überraschen, denn in der *Cognitive Load Theory* (z. B. Paas et al. 2003) gilt die Elementinteraktivität ausdrücklich als Maß intrinsischer kognitiver Belastung (in Interaktion mit Vorwissen und Arbeitsgedächtniskapazität der Lernenden). Die von Karpicke und Aue (2015) ins Spiel gebrachten Komplexitätsmetriken beziehen sich allerdings eindeutig auf die Textkomplexität. Entsprechend bliebe Karpicke und Aues (2015) Kritik einer mangelnden Objektivität der Erfassung der Elementinteraktivität bei van Gog und Sweller (2015) zwar gerechtfertigt, ihre Vorschläge zu einer objektiven Erfassung gingen aber am eigentlichen Konzept der Elementinteraktivität vorbei.

In einer späteren Version der *Cognitive Load Theory* ist die Ausgangssituation allerdings weniger klar bzw. sogar eine explizit andere. Sweller (2010, S. 125) schlägt vor, Elementinteraktivität nicht allein auf intrinsische Belastung, sondern auch auf extrinsische Belastung zu beziehen: „I suggest that element interactivity is the major source of working memory load underlying extraneous as well as intrinsic cognitive load. If element interactivity can be reduced without altering what is learned, the load is extraneous; if element interactivity only can be altered by altering what is learned, the load is intrinsic". Entsprechend dieser Konzeptualisierung könnte sich eine hohe Elementinteraktivität also sowohl aus den Inhalten als auch aus deren Darstellung ergeben und die von Karpicke und Aue (2015) genutzten Textkomplexitätsmetriken wie *Flesch reading ease* (Flesch 1948) wären eine Möglichkeit, Letzteres objektiv zu erfassen. Es erscheint uns allerdings äußerst fraglich, ob eine hohe Elementinteraktivität in diesem Sinn die von van Gog und Sweller (2015) angeführten Folgen für den Nutzen von Abrufübung hat. Das Argument von van Gog und Sweller (2015) bezieht sich darauf, dass Materialien hoher Komplexität es ohnehin erfordern, viele Verbindungen zwischen verschiedenen Elementen zu ziehen, was eher der Elementinteraktivität als Quelle intrinsischer Belastung entspricht und damit der Komplexität des Inhalts. In diesem Fall wären die von Karpicke und Aue (2015) ins Spiel gebrachten Textkomplexitätsmetriken dem Vorgehen von van Gog und Sweller (2015) zwar in puncto Objektivität überlegen, die fehlende Validität macht diesen Vorteil aber zunichte. Auch der Einsatz von *Coh-Metrix* (Graesser et al. 2014) zur Erfassung der lokalen Textkohäsion, was laut Karpicke und Aue (2015) dem Konzept der Elementinteraktivität am nächsten kommt, ist wenig überzeugend. So lässt sich anhand von Argumentüberlappung zwar erfassen, wie viele explizite referentielle Bezüge in den Texten enthalten sind, aber nicht, ob es für das Verstehen *erforderlich* ist, mehrere Konzepte gleichzeitig im Arbeitsgedächtnis verfügbar zu halten. Zudem sind lokal kohärentere Texte eher leichter verständlich (z. B. Burkhart et al. 2020; Lachner et al. 2017).

Ein weiterer zentraler und aus unserer Sicht gravierender Kritikpunkt von Karpicke und Aue (2015) an den Schlussfolgerungen von van Gog und Sweller (2015) war, dass keines der berichteten Experimente tatsächlich Komplexität oder Elementinteraktivität manipulierte, weder im Hinblick auf lernstoffbezogene noch im Hinblick auf textbezogene Aspekte. Stattdessen enthielten die Experimente nur solche Materialien, von denen angenommen wurde, dass sie komplex (oder weniger komplex) sind.

Auf Basis der bisherigen Datengrundlage ist es daher schwer zu ergründen, ob Komplexität bzw. hohe Elementinteraktivität des Lernmaterials tatsächlich eine Randbedingung für das Auftreten eines Testungseffekt darstellt. Wie bereits dargelegt, ist ein wesentlicher Grund hierfür die Unklarheit darüber, wie die Komplexität von Lernmaterialien möglichst objektiv erfasst werden kann. Dieses Problem ist nur dann einigermaßen gut gelöst, wenn sich die Erfassung der Komplexität ausschließlich auf die sprachliche Komplexität von Texten bezieht. Wenn es jedoch um die Elementinteraktivität des Lern*inhalts* geht, sieht die Situation anders aus. Hier fehlen bislang überzeugende, objektive und reproduzierbare Maße, die über eine verbale Beschreibung der miteinander in Verbindung zu bringenden Elemente hinausgehen. Eine der Ursachen für das Fehlen einer entsprechenden Metrik ist, dass bereits die Bestimmung eines Elements subjektive Aspekte enthält. So wird als typisches Beispiel für Lernmaterial, bei dem nur isolierte Elemente gelernt werden müssen, das Vokalbellernen genannt (Sweller 2010). Ein Vokabelpaar wird also als ein Element angesehen. Aus unserer Perspektive ist allerdings mindestens ebenso plausibel anzunehmen, dass das Vokabellernen die Verbindung von zwei Elementen erfordert – dem muttersprachlichen und dem fremdsprachlichen Begriff – und erst wenn diese Verbindung gelernt wurde, kann das Vokabelpaar gespeichert werden.

Allein dieser Punkt veranschaulicht, wie schwer es ist, die Anzahl von Elementen zu bestimmen, die gleichzeitig im Arbeitsgedächtnis gehalten werden müssen. Zudem muss bestimmt werden, ob die Gesamtheit der Elemente gleichzeitig verarbeitet und miteinander in Verbindung gebracht werden muss, um einen Text zu verstehen, oder ob ein Verständnis auch erreicht werden kann, indem einzelne Teilkomplexe zunächst separat und damit konsekutiv verarbeitet und verstanden werden. In jedem Fall beinhaltet eine Entscheidung über die Anzahl der zeitgleich zu repräsentierenden Elemente große subjektive Spielräume. Insgesamt macht dies deutlich, dass der Begriff Elementinteraktivität eine Präzision nahelegt, die die entsprechenden Operationalisierungsversuche nicht einlösen können (s. hierzu auch Karpicke und Aue 2015).

6 Eine Alternativerklärung für den Einfluss von Komplexität auf den Testungseffekt

Neben diesen messtheoretischen Problemen besteht unserer Auffassung nach allerdings noch ein weiteres Problem: Selbst, wenn man davon ausgeht, dass die als komplex eingestuften Materialien in Studien, die keinen Testungseffekt zeigten, tatsächlich komplex bzw. von hoher Elementinteraktivität waren, gibt es eine Alternativerklärung für diese Befunde. Komplexe Inhalte sind schwerer verständlich als weniger komplexe und erfordern daher in der Regel ein besonders aufmerksames Lesen oder ein erneutes Lesen kritischer Textpassagen und damit typischerweise eine längere Lernzeit. Es ist daher möglich, dass die initiale Lernzeit in den Studien mit komplexen Materialien schlicht nicht ausreichend war, um ein solides Verständnis zu gewährleisten. In diesem Fall könnte es durchaus sein, dass die Lernenden zu dem Zeitpunkt, zu dem die Wiederholungsphasen begannen, noch nicht über eine ausreichend gute Wissensbasis verfügten. Der Abruf in den Übungstests wäre dann

unvollständig, und es könnte sogar sein, dass einige der Übungsfragen auf der Basis von Fehlkonzepten falsch beantwortet wurden. Fehler bzw. Fehlkonzepte würden unter diesen Umständen konsolidiert, vor allem dann, wenn es keine Möglichkeit gibt, die Lernmaterialien erneut anzuschauen oder in anderer Form Feedback zu erhalten. Insbesondere der Nutzen des direkten Testungseffekts sollte so eingeschränkt sein, aber auch indirekte Testungseffekte (bzw. Effekte des testpotenzierten Lernens) sollten hier eher gering ausfallen. So ist davon auszugehen, dass Feedback (etwa durch die erneute Vorlage eines Textes) bei mangelndem Verständnis weniger gezielt genutzt werden kann, als wenn nur einzelne Wissenslücken geschlossen oder überschaubar wenige und klar benennbare Verständnisprobleme abgeklärt werden müssen.

In klassischen Experimenten zum Testungseffekt geht der Konsolidierungsphase, in der typischerweise eine Lese- und eine Testungsphase verglichen werden, eine einzelne Lesephase voraus. Bei komplexen Texten könnte eine Wiederholung in Form einer Verlängerung der Lesephase als Verlängerung der initialen Lernphase genutzt werden, um das zunächst mangelnde Verständnis nach der ersten Lernphase zu kompensieren und so ein Grundverständnis sicherzustellen. Im Falle einfacher Materialien wäre eine solche Verlängerung nicht oder kaum erforderlich, und das erneute Lesen würde allein der Konsolidierung des Gelernten dienen. Da Testung allerdings ein effektiveres Mittel der Konsolidierung darstellt als passives Lesen, käme es in diesem Fall zu einem Testungseffekt. Bei komplexen Lernmaterialien würde die bessere Konsolidierung durch Testung hingegen gerade ausreichen, um den Nachteil durch das initial schlechtere Verständnis auszugleichen. Dieser Argumentation folgend wäre es allerdings naheliegend anzunehmen, dass Testung auch bei komplexem Material effektiv sein kann, wenn die Lernenden bis zum Beginn der Wiederholungsphasen ein hinreichend gutes Verständnis des Lernstoffs erlangt haben und so mit dem Testen und erneuten Lesen (oder anderen Bedingungen, in denen das Lernmaterial weiter bzw. erneut verfügbar ist) tatsächlich zwei Wiederholungsgelegenheiten miteinander verglichen werden. Sollte dies der Fall sein, so wäre eher das Textverständnis als die Komplexität eine Randbedingung für den Testungseffekt. Dies vorausgesetzt, würde sich nicht die Frage stellen, *ob* man bei komplexen Lernmaterialien auf Testung verzichten sollte, sondern wann mit dem Testen begonnen werden sollte und wie initiale Lernbedingungen gestaltet werden sollten, um die Vorteile der Testung auch bei komplexen Materialien zu ermöglichen.

Der eingangs beschriebene Befund, dass der langfristige Nutzen der Testung mit der Anzahl der Übungstests steigt, legt nahe, dass für einen dauerhaften Nutzen die Abrufübung so früh wie möglich in den Lernprozess einbezogen werden sollte, damit das Wissen in möglichst vielen (erfolgreichen) Abrufversuchen gefestigt werden kann. Die Überlegungen zum Testen mit komplexen Materialien lassen hier allerdings die Frage aufkommen, wann dieser früheste Zeitpunkt ist. Aus dem Umstand, dass eine angemessene Wissensbasis eine notwendige Voraussetzung dafür ist, dass der Abruf der entsprechenden Informationen gelingen und das Wissen konsolidiert werden kann, ergibt sich, dass nicht zu früh mit der Abrufübung begonnen werden sollte. Wenn ein Sachverhalt noch nicht ausreichend verstanden wurde, kann nichts Substanzielles gefestigt werden oder es könnten sogar Fehlkonzepte gefestigt werden, vor allem dann, wenn es kein Feedback gab, anhand dessen die entsprechenden

Fehlkonzepte korrigiert werden konnten (für ein ähnliches Argument s. Roelle und Nückles 2019; Rummer 2021). Für die Frage, wann man im Verlauf des Lernprozesses mit dem Testen beginnen sollte, gilt es also, eine Balance zu finden zwischen „so früh wie möglich", um mehrere Gelegenheiten zum Üben des Abrufs zu ermöglichen (und den Abruf „schwer" zu gestalten), und „nicht zu früh", um zu vermeiden, dass konsolidiert wird, bevor ausreichend Wissen aufgebaut worden ist.[2]

Eine einfache Möglichkeit, die Lernbedingungen so zu verändern, dass sie sowohl ausreichende Lerngelegenheiten als auch optimale Übungsmöglichkeiten für den Abruf ermöglichen, könnte entsprechend darin bestehen, die initiale Lernphase bei komplexen Lernmaterialien länger zu gestalten als bei weniger komplexen Inhalten (s. auch Rummer 2021). Alternativ könnten die Lernenden bereits vor der ersten Lernphase besser auf das Verständnis komplexer Materialien vorbereitet werden, etwa im Sinne eines „pre-training" (z. B. Mayer et al. 2002). Hier werden den Lernenden die Eigenschaften zentraler Konzepte vermittelt, bevor die eigentliche Instruktion beginnt. Dies könnte auch das Verständnis komplexer Sachverhalte in der Lernphase verbessern, so dass die Lernenden von der Abrufübung profitieren können.

Allerdings ist es an dieser Stelle wichtig, darauf hinzuweisen, dass diese Überlegungen zur Rolle der Komplexität in Bezug auf den Testungseffekt bislang spekulativ sind und noch nicht empirisch überprüft wurden. Hier wäre es erforderlich, sowohl die Komplexität des Lernmaterials zu manipulieren (unter Berücksichtigung der bereits beschriebenen Schwierigkeiten) als auch die Gestaltung der initialen Lernphase(n). Sollten sich in solchen Studien unsere Annahmen bestätigen, wäre nicht die Komplexität oder Elementinteraktivität des Materials ausschlaggebend für das (Nicht-)Auftreten eines Testungseffekts, sondern das erreichte Verständnis zum Zeitpunkt der Konsolidierungsphase, das wiederum von der Komplexität des Lernmaterials, dem Vorwissen der Lernenden und der ihnen zur Verfügung stehenden initialen Lernzeit abhängt. Insgesamt lässt sich jedoch festhalten, dass zurzeit sowohl aufgrund der bislang unzureichenden Befundlage als auch aufgrund der Unklarheit bezüglich der Frage, was die Komplexität von Lernmaterialien ausmacht, Komplexität nicht als etablierte Randbedingung für das Auftreten eines Testungseffekts gelten kann.

[2] In diesem Kontext ist eine weitere Forschungslinie erwähnenswert, die nahelegt, dass auch Fragen, die maximal früh, und zwar bereits vor der ersten Beschäftigung mit dem Lernmaterial gestellt werden, die Lernleistung positiv beeinflussen können (z. B. Carpenter und Toftness 2017; Kornell et al. 2009). Laut Arnold und McDermott (2013) ist in diesem Fall allerdings eher von Generierung als von Testung zu sprechen, da hier kein Abruf von zuvor enkodierten Informationen stattfindet. Zudem nutzten die bisherigen Studien typischerweise Lernmaterial, das als von geringer Elementinteraktivität einzustufen wäre (v. a. Wortpaare und abseitige Wissensfragen, z. B. Kornell et al. 2009; Kornell 2014). In ersten Studien mit komplexeren Materialien wie Essays (Richland et al. 2009) und Lehrvideos (Carpenter und Toftness 2017; Toftness et al. 2018) zeigten sich ebenfalls Vorteile durch vorab gestellte Fragen – die Fragen adressierten allerdings jeweils isolierte Fakten. Ein möglicher Einfluss der inhaltlichen Komplexität ist für vorab gestellte Fragen zurzeit also noch schwerer zu beurteilen als für nach dem initialen Lernen gestellte Fragen.

7 *Open-Book*-Testung als Möglichkeit der Flexibilisierung des Lernens

Ein vielversprechender Ansatz, um einerseits das Verständnis zu fördern und andererseits bereits früh Abrufübung zu ermöglichen, könnte die Verwendung von *Open-Book*-Tests anstelle der typischerweise verwendeten *Closed-Book*-Tests sein. In *Open-Book*-Tests müssen die Lernenden wie in einer typischen Testungsbedingung Übungsfragen beantworten, können dazu aber die Lernmaterialien, also Texte oder eventuell gemachte Notizen, zu Rate ziehen. Wenn Lernende mit einer Frage konfrontiert werden, die sie nicht unmittelbar beantworten können, bietet ein *Open-Book*-Test die Möglichkeit, bestimmte Textpassagen erneut zu studieren und das anfängliche Verständnis zu verbessern, um dann dazu in der Lage zu sein, die Frage zu beantworten. Dies sollte besonders dann hilfreich sein, wenn es sich um komplexe Lernmaterialien handelt, die die Lernenden nach einer initialen Lesephase noch nicht ausreichend verstanden haben. In diesem Fall böte, wie bereits für das wiederholte Lesen argumentiert, die Verfügbarkeit des Textes die Möglichkeit, das unzureichende Textverständnis zu kompensieren. Anders als das rein wiederholende Lesen regt *Open-Book*-Testung jedoch trotzdem zur Abrufübung an. Die Beantwortung von Übungsfragen mit verfügbaren Materialien sollte allerdings im Vergleich zu einem *Closed-Book*-Test das Ausmaß an Abrufübung reduzieren, wenn – wie gerade ausgeführt – zumindest einige Informationen nicht aus dem Gedächtnis abgerufen, sondern im Text nachgeschlagen werden (für ein ähnliches Argument s. Roelle und Nückles in diesem Themenheft; Rummer et al. 2019).

Bei *Open-Book*-Tests müssen die Lernenden also eine Balance zwischen erneutem Lesen und aktivem Abrufen herstellen. Ideal aus Sicht der Testungsliteratur wäre, wenn sich die Lernenden dabei darum bemühen würden, zunächst so viele Informationen wie möglich aus dem Gedächtnis abzurufen und von der Möglichkeit der Materialeinsicht nur dann Gebrauch zu machen, wenn sie bestimmte Informationen, auch mit viel Anstrengung, nicht abrufen können. In diesem Fall sollten sie sowohl in stärkerem Maße vom direkten Testungseffekt profitieren als auch, wenn sie sich erneut dem Text zuwenden, vom indirekten Testungseffekt. Darüber hinaus kann der verfügbare Text genutzt werden, um die abgerufenen Antworten hinsichtlich ihrer Korrektheit zu überprüfen oder Antworten zu präzisieren. Voraussetzung für ein solches Vorgehen wäre allerdings das Wissen um die prinzipielle Nützlichkeit von Abrufübung. Dass Lernende tatsächlich eine solche Strategie verwenden, ist aus unserer Sicht eher unwahrscheinlich (s. auch Waldeyer et al. 2020). Unserer Einschätzung nach ist es wahrscheinlicher, Lernende beantworten allenfalls die Fragen, deren Beantwortung ihnen besonders leichtfällt, ohne Einsicht in das Lernmaterial zu nehmen, und alle anderen Fragen mehr oder weniger unter Einsichtnahme in die Lerntexte. Allerdings sind dies Spekulationen, die einer weiteren experimentellen Prüfung unterzogen werden müssten.

Eine Möglichkeit, den Lernprozess effektiver zu gestalten, bestünde nach dem bisher Gesagten darin, die Lernenden ausdrücklich dazu aufzufordern, zunächst alle Fragen zu beantworten, ohne die Materialien einzusehen, und die Texte erst dann zu nutzen, wenn die Fragen nicht ohne erneute Materialeinsicht beantwortet werden können. Auf diese Art und Weise könnten die Lernenden sowohl vom direkten als auch vom indirekten Testungseffekt profitieren. Ähnlich gingen Waldeyer et al.

(2020) in einer Bedingung vor, die den Lernenden den flexiblen Wechsel zwischen einer *Closed-Book*- und einer *Open-Book*-Bearbeitung ermöglichte. Der Text war hier standardmäßig nicht sichtbar, konnte aber von den Lernenden sichtbar gemacht werden und verschwand dann wieder, entweder auf Initiative der Lernenden oder automatisch nach 30 s. Im Vergleich zu einer reinen *Open-Book*- und einer reinen *Closed-Book*-Bearbeitung führte diese Bedingung zum größten Lernerfolg, während sich die anderen Bedingungen nicht unterschieden.

8 Zusammenfassung und Schlussfolgerung für die Praxis

Zusammenfassend lässt sich feststellen, dass der Testungseffekt durchaus robust ist und für unterschiedliche Testungsverfahren und im Vergleich zu unterschiedlichen Kontrollgruppen sowohl im Labor als auch im Feld nachgewiesen werden konnte. Allerdings ist es eine vieldiskutierte und noch immer offene Frage, ob sich der Testungseffekt auch unter Verwendung komplexer Lernmaterialien zuverlässig zeigt. Laut van Gog und Sweller (2015) sollte dies nicht der Fall sein, da bei der Verarbeitung von Materialien mit hoher Elementinteraktivität auch ohne Abrufübung zahlreiche Verbindungen zwischen den einzelnen Elementen hergestellt werden (müssen) und so kein zusätzlicher Nutzen durch Testung zu erwarten ist. Sie verweisen auf eine Reihe von Experimenten, die diese Hypothese unterstützen und bei komplexen Materialien keinen Testungseffekt zeigen. Allerdings wurde die von van Gog und Sweller (2015) vorgenommene subjektive Kategorisierung der Materialkomplexität vehement kritisiert (Karpicke und Aue 2015). Wir haben dafür argumentiert, dass diese Diskussion zum einen durch unausgesprochene Unterschiede in der Interpretation des Konzepts Komplexität bzw. Elementinteraktivität erschwert wird und dass nicht die Komplexität selbst für das Ausbleiben von Testungseffekten verantwortlich ist, sondern der damit einhergehende Umstand, dass die Inhalte zum Zeitpunkt der Testung noch nicht ausreichend gut verstanden waren. Sollte sich diese Annahme bestätigen, ergäben sich daraus andere praktische Implikationen als aus der Annahme, Komplexität sei eine Randbedingung für den Testungseffekt. In diesem Fall wäre es ratsam, bei komplexen Materialien nicht auf Testung zu verzichten, sondern mittels ausreichender Lesezeit und anderen Methoden der Verständnissicherung (z. B. Pre-Training zur Einführung von Schlüsselbegriffen) dafür zu sorgen, dass zum Zeitpunkt der Testung die Inhalte weitgehend verstanden sind. In diesem Kontext könnten bei komplexen Inhalten auch *Open-Book*-Testungen hilfreich sein, die es den Lernenden erlauben, selbständig aus einem Abrufmodus wieder in einen Lesemodus zu wechseln, wenn sie auf Verständnisprobleme stoßen. Ein effektiver Einsatz setzt allerdings sowohl das metakognitive Wissen über den Nutzen von Abrufübung als auch gute metakognitive Monitoring- und Regulationsfähigkeiten voraus, um Verständnisprobleme zu entdecken und entsprechend darauf zu reagieren (s. auch Roelle und Renkl 2020).

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Abott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, *11*(1), 159–177. https://doi.org/10.1037/h0093018.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199.

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2016). Ecological validity of the testing effect. *Teaching of Psychology*, *44*, 18–23. https://doi.org/10.1177/0098628316677492.

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist, C., & Jonsson, B. (2021). Retrieval practice: beneficial for all students or moderated by individual differences? *Psychology Learning & Teaching*, *20*(1), 21–39. https://doi.org/10.1177/1475725720973494.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Hrsg.), *Information processing and cognition: the Loyola Symposium* (S. 123–144). : Lawrence Erlbaum.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*, 59–68.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. https://doi.org/10.1037/a0035934.

Burkhart, C., Lachner, A., & Nückles, M. (2020). Using spatial contiguity and signaling to optimize visual feedback on students' written explanations. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000607.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1563–1569. https://doi.org/10.1037/a0017021.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhnces subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405.

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory & Cognition*, *6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014.

Chan, J. C., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: a theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *40*, 104.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*, 2412. https://doi.org/10.3389/fpsyg.2018.02412.
Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3.
Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327.
Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 106–114. https://doi.org/10.1037/a0033699.
Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729.
Kühn, A. (1914). Über Einprägung durch Lesen und durch Rezitieren. [About memorisation through reading and recitation]. *Zeitschrift für Psychologie*, *68*, 396–481.
Lachner, A., Burkhart, C., & Nückles, M. (2017). Mind the gap! Automated concept map feedback supports students in writing cohesive explanations. *Journal of Experimental Psychology: Applied*, *23*(1), 29–46. https://doi.org/10.1037/xap0000111.
Mayer, R. E., Mathias, A., & Wetzell, K. (2002). Fostering understanding of multimedia messages through pre-training: evidence for a two-stage theory of mental model construction. *Journal of Experimental Psychology: Applied*, *8*(3), 147–154. https://doi.org/10.1037/1076-898X.8.3.147.
McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: effective and portable. *Psychological Science*, *20*(4), 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9.
Nguyen, K., & McDaniel, M. A. (2016). The JOis of text comprehension: supplementing retrieval practice to enhance inference performance. *Journal of Experimental Psychology: Applied*, *22*(1), 59–71. https://doi.org/10.1037/xap0000066.
Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1.
Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710. https://doi.org/10.1037/bul0000151.
Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning*. IES Practice Guide. NCER 2007–2004. : National Center for Education Research.
Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology*, *5*, 1–5. https://doi.org/10.3389/fpsyg.2014.00286.
Rawson, K. A. (2015). The status of the testing effect for complex materials: still a winner. *Educational Psychology Review*, *27*(2), 327–331. https://doi.org/10.1007/s10648-015-9308-4.
Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496.
Roediger III, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.
Roediger III, H. L., & Karpicke, J. D. (2006b). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x.
Roelle, J., & Nückles, M. (2019). Generative learning versus retrieval practice in learning from text: the cohesion and elaboration of the text matters. *Journal of Educational Psychology*, *111*, 1341–1361. https://doi.org/10.1037/edu0000345.
Roelle, J., & Renkl, A. (2020). Does an option to review instructional explanations enhance examplebased learning? It depends on learners' academic self-concept. *Journal of Educational Psychology*, *112*, 131–147. https://doi.org/10.1037/edu0000365.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559.

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review*, *21*(6), 1516–1523. https://doi.org/10.3758/s13423-014-0625-2.

Rummer, R. (2021). Der Testungseffekt beim Lernen mit Texten: Ein Beispiel für das schwierige Verhältnis zwischen Grundlagenforschung und Anwendung. *Psychologische Rundschau*, *72*(4), 259–272. https://doi.org/10.1026/0033-3042/a000518.

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, *23*(3), 293–300. https://doi.org/10.1037/xap0000134.

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: a field experiment. *Frontiers in Educational Psychology*, *10*, 463. https://doi.org/10.3389/fpsyg.2019.00463.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: a meta-analytic perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138.

Tempel, T., & Frings, C. (2019). Testing enhances motor practice. *Memory & cognition*, *47*(7), 1270–1283. https://doi.org/10.3758/s13421-019-00932-6.

Tempel, T., & Pastötter, B. (2021). Abrufeffekte im Gedächtnis: Ein Überblick zur aktuellen Grundlagenforschung. [Retrieval effects in memory: an overview of current basic research]. *Psychologische Rundschau*, *72*(4), 249–258. https://doi.org/10.1026/0033-3042/a000517.

Toftness, A. R., Carpenter, S. K., Lauber, S., & Mickes, L. (2018). The limited effects of prequestions on learning from authentic lecture videos. *Journal of Applied Research in Memory and Cognition*, *7*(3), 370–378. https://doi.org/10.1016/j.jarmac.2018.06.003.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*, 247–264. https://doi.org/10.1007/s10648-015-9310-x.

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001.

Wirtz, M. (2013). *Dorsch – Lexikon der Psychologie* (16. Aufl.). Bern: Huber.

Witasek, S. (1907). Über Lesen und Rezitieren in ihren Beziehungen zum Gedächtnis. [On reading and recitation in their relations to memory]. *Zeitschrift für Psychologie*, *44*, 161–185.

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: a review of the forward testing effect. *npj: Science of Learning*. https://doi.org/10.1038/s41539-018-0024-y.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*. https://doi.org/10.1037/bul0000309.