journal für lehrerInnenbildung
jlb
no.1
2022

Testen und Bewerten

Bibliografie:
Bernhard Hauser:
Testing Drives Learning?
journal für lehrerInnenbildung, 22 (1), 78-87.
https://doi.org/10.35468/jlb-01-2022-07

Gesamtausgabe online unter:
http://www.jlb-journallehrerinnenbildung.net
https://doi.org/10.35468/jlb-01-2022

ISSN 2629-4982

klinkhardt

07

Bernhard Hauser

Testing Drives Learning?

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07

Tests beeinflussen den Alltag von Schule und Hochschule auf mannigfaltige Weise – auch in der Aus- und Weiterbildung von Lehrpersonen. In diesem Beitrag werden die Begriffe „vergleichende Tests mit Rankings", „Teaching-to-the-Test" und „Testing-Effekt" erläutert und deren Wirkungen mit empirischen Befunden eingeordnet.
Vergleichstests mit Rankings, mit denen Schulen vor Ort mit anderen Regionen und Ländern verglichen werden, stoßen bei Lehrpersonen mehrheitlich eher auf Ablehnung und haben auch eher ungünstigen Einfluss auf deren Belastungsempfinden und Berufszufriedenheit (Reid, 2011; Smith & Holloway, 2020). Dies, obwohl derartige Tests als Instrumente der Qualitätssicherung und Bildungssteuerung die Leistungsfähigkeit von Schulen sicherstellen und auf diese eher positiv wirken, indem sie beispielsweise oft Grundlage sind für nachfolgende Reformen. Teaching-to-the-Test und damit das auf einen kommenden Test ausgerichtete Lernen hat Vor- und Nachteile, die stark mit der Beschaffenheit des Tests zusammenhängen. Ist der Test gut, dann hat das auf den Test ausgerichtete Lernen vor allem Vorteile. Der Einsatz von Tests während dem Lernprozess hat positive Effekte (Testing-Effekt) und kann das Lernen auf unterschiedliche Weise verbessern. Zum einen bei anspruchsvollen selektionsrelevanten und/oder qualifizierenden Tests, deren Muster bekannt sind, und die damit das vorangehende Lernen und Lehren beeinflussen, die als Learning-to-the-Test und als Teaching-to-the-Test positiv oder negativ wirken können. Insbesondere nicht-selektionsrelevante sogenannte formative Tests können das Lernen fördern, indem sie funktional als Teil einer Feedback-Kultur innerhalb transparenter Lernziele fungieren.

Vergleichende Tests mit Rankings

Im Zentrum der Kritik steht vor allem das High-stakes-Testing. Dabei handelt es sich um zentrale, internationale oder interregionale Leistungstests, die außerhalb der lokalen Schule administriert werden und deren Ergebnisse beträchtlichen Einfluss auf das unmittelbare Umfeld der überprüften Klassen haben: So fließen in der Schweiz die individuellen Ergebnisse überregionaler Vergleichstests wie Klassencockpit, einem geeichten Instrument zur Standortbestimmung der Leistungen in Mathematik und Sprache der Volksschule, in die Berechnung von Schlussnoten von Lernenden der Volksschule ein. Sie können damit

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07

das schulische Weiterkommen der Lernenden in nachfolgende Klassen beeinflussen. In einigen Ländern haben die Ergebnisse von Vergleichstests sogar Einfluss auf Festanstellung oder Kündigung beziehungsweise Weiterbeschäftigung von Lehrpersonen. Zuweilen finden sich auch direkte Folgen für die Schulen: Diese können im nationalen oder regionalen öffentlichen Diskurs unter Druck geraten nach einer als zu wenig gut eingeschätzten Position im interregionalen oder internationalen Wettbewerb. Rangpositionen von Schulen können – je nach Land – auch einschneidende Maßnahmen nach sich ziehen in den Bereichen Finanzierung, Akkreditierung, bis hin zur Schliessung von Schulen.

Diese Vielfalt an Wirkungen findet auch ihren Niederschlag in Einstellungen und Wahrnehmungen von Lehrpersonen. So glauben angehende US-Amerikanische Lehrpersonen mit überwältigender Mehrheit, dass zu oft getestet würde und dass der Fokus auf das erfolgreiche Bestehen der Tests zu viel Einfluss auf den Unterricht habe (Reid, 2011). Es sind vor allem persönliche Erfahrungen, welche diese Skepsis gegenüber dem Testen alimentieren. Diese Studierenden des Lehramts äußern sich auch gegen ein Teaching-to-the-Test. So müsse etwa im Mathematikunterricht das konzeptuelle Verstehen im Zentrum stehen, und weniger das zu wenig verstehensorientierte Automatisieren mathematischer Regeln und Prozeduren. Letzteres werde jedoch durch Teaching-to-the-Test leider befördert. Auch bei Lehrpersonen und Lehramtsstudierenden in deutschsprachigen Ländern ist eine Skepsis diesen Tests gegenüber nicht unbekannt.

Vergleichstests dieser Art haben auch Folgen auf die Psyche von Lehrpersonen. So führen internationale Vergleichstests wie PISA zu einem zunehmenden Belastungsempfinden bei Lehrpersonen, wie Smith und Holloway (2020) in einem Überblick zu Befunden aus 33 Ländern seit dem Jahr 2013 zeigen konnten. Diese Belastungen lassen sich unter anderem darauf zurückführen, dass Lehrpersonen sich verantwortlich fühlen für das bildungsmäßige Abschneiden des eigenen Landes oder der eigenen Region. Dabei zeigte sich: je höher die Intensität der Testkultur, desto stärker waren die (negativen) Wirkungen auf die Berufszufriedenheit von Lehrpersonen. So hatte beispielsweise das Abschneiden des Landes Fürstentum Liechtenstein in der ersten PISA-Studie aus dem Jahr 2000 zu einem ganzseitigen Inserat in der wichtigsten Zeitung des Landes geführt. Darin bemängelten die einheimischen Wirtschaftsverbände das im Vergleich zu anderen eu-

ropäischen Ländern schlechte Abschneiden der Schulen des eigenen Landes mit dem Titel *Jetzt muss gehandelt werden*. Zu diesem Zeitpunkt führte ich eine schulinterne Weiterbildung in einem Dorf des Fürstentums durch und erinnere mich noch gut an die emotional und heftig geführte Debatte im dortigen Lehrpersonenzimmer. Derart öffentlich geführte Kritik löst bei Lehrpersonen viel aus, wie die Debatte um den Lehrberuf zwischen der Hochstilisierung von Lehrpersonen als Hoffnungsträger und deren Verachtung als Versager (Bastian & Combe, 2007) zeigt. Auch wenn sich Lehrpersonen nach mehr als 20 Jahren PISA inzwischen einigermassen damit arrangiert haben, geht diese Form von öffentlichem Rechtfertigungsdruck von Bildungsinvestitionen auch heute noch nicht spurlos an ihnen vorbei. Die Folgen des Trends zu vermehrtem Wettbewerb zwischen Nationen und Regionen wurden wohl unterschätzt. Die zuweilen auf derartige Vergleichstestungen erfolgenden Bildungsreformen mögen oft zu Verbesserungen des Systems führen, beeinträchtigen aber nicht selten die Zufriedenheit von Lehrpersonen, was sich in einem Ansteigen von Stress und Burnout ausdrückt (Smith & Holloway, 2020). So führen zentrale Leistungs- und Vergleichstests bei Novizen-Lehrpersonen vermehrt zu lehrpersonen-zentrierten Lektionen (Ro, 2019), weil diese sich bei diesem Vorgehen in der Vorbereitung auf diese Tests sicherer zu fühlen. Dieses Verhalten lässt sich auch als Teaching-to-the-Test charakterisieren, eine Unterrichtsphilosophie, die oft vorschnell als dem Lernen abträglich beschrieben wird (vgl. unten).

Teaching-to-the-Test

Begegne ich früheren Kolleg*innen aus der eigenen Schulzeit, oder Kommiliton*innen aus dem Studium an der Universität, dann drehen sich unsere Gespräche nicht selten um Erinnerungen an Lehrer*innen, an Professor*innen, und um deren Ruf, der in engem Zusammenhang mit deren Anforderungen in den Prüfungen stand. Je breiter und tiefer das in deren Prüfungen eingeforderte Wissen und Können war, desto mehr wurde gelernt. Ähnliches berichten auch Studierende von heute, vor allem ehemalige. Diese – basierend auf ausgewählten subjektiven Theorien erfolgreichen – Pädagog*innen unterschieden sich aber nicht nur in der Anspruchshöhe ihrer Prüfungen, sie unterschieden sich auch darin, wie transparent sie die in den Prüfungen geforder-

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07

ten Kompetenzen kommunizierten und wie gut sie Lernende darauf vorbereiteten. Das Lehren für die Prüfung – auch Teaching-to-the-Test genannt – ist also nichts Neues.

Teaching-to-the-Test beginnt bei Learning-to-the-Test. Vielen Lehrenden aller Stufen ist die Frage *Kommt das an der Prüfung?* sehr vertraut. Diese Orientierung Lernender und Studierender am Inhalt, den sie für eine kommende Prüfung für bedeutsam erachten, wird auch als Washback-Effekt (nach Cheng & Curtis, 2003) beschrieben. Es geht um den Einfluss, den Tests auf Lehre und Lernen haben (Alderson & Wall, 1993). „Was gemessen wird erhält Bedeutung und Wert, was wiederum die Lehre beeinflusst" (Mc Ewan, 1995, S. 42; Übersetzung durch den Autor). Viele Lehrende, auch an Hochschulen, weisen bei sehr wichtigen Inhalten und Zusammenhängen nicht selten darauf hin, dass diese später auch Gegenstand der Prüfung sein werden. Im Kern geht es darum, wie stark die kommende – in der Regel benotete – Prüfung auf das vorangehende Lernen einen Einfluss hat. Dieser Effekt kann negativ sein, wenn der Unterricht sehr eng nur auf das Bestehen der Tests fokussiert ist, z. B. durch mechanisches Üben von im Test vorkommenden Aufgabentypen. Allerdings setzt dies einen eher schlechten Test voraus. Bei qualitativ hochwertigen Tests, wie dies bei den meisten PISA-Testaufgaben der Fall ist, kann Teaching-to-the-Test sehr ertragreich sein. Auf diesen positiven Aspekt, der vermutlich bedeutsamer ist als die häufig beklagten Nachteile, wies schon Andreas Helmke hin:

> „Versteht man hingegen ‚teaching-to-the-test' so, dass anlässlich einer Vergleichsarbeit anspruchsvolle Aufgabentypen verstärkt thematisiert werden, und dass dabei ausreichende Gelegenheiten für horizontalen (andere Kontexte) und vertikalen (höhere Komplexität, neue Fragestellungen) Transfer gegeben werden, dann würde ich das als eine intelligente Form des Übens betrachten." (Helmke, 2007, 62f.).

Wichtig ist, dass relevante und anspruchsvolle Kompetenzen Teil der Prüfung sind, sonst werden diese zu wenig gelernt. Wird zum Beispiel bei Studierenden in schriftlichen Arbeiten die Kompetenz des korrekten Zitierens höher gewichtet als das Durchdringen bedeutsamer und evidenzbasierter Theorien und deren Anwendung, dann wäre das eine Folge von zu banalen Bestehenskriterien und damit ein Negativ-Beispiel für Teaching-to-the-Test. Teaching-to-the-Test ist nur dann ein Problem, wenn der Test schlecht ist.

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07

Der Testing-Effekt

Eine besondere Variante von Teaching-to-the-Test ist der Einsatz von Übungstests. Damit können Teilbereiche einer nachfolgenden Prüfung vorher geübt werden, ohne dass das Ergebnis Konsequenzen hat für das Bestehen und Weiterkommen in der Schule. Übungstests erhöhen die Transparenz und geben Feedback zum aktuell erreichten Niveau auf dem Weg zum Können. Hierzu gibt es interessante Befunde, vor allem aus dem Bereich der Gesundheitsberufe. Auch wenn in diesem Berufsfeld Fertigkeiten und Kompetenzen oft klarer ab- und eingegrenzt werden können als Kompetenzen im Lehrberuf, so liefern Befunde dieser Studien durchaus interessante Anregungen für den Erwerb des Lehrberufs.

Der verbesserte Lernertrag durch den Einsatz von Tests zu Übungszwecken wird als Testing-Effekt bezeichnet. Danach stellen sich Lernfortschritte ein, wenn Lernende einen Übungstest zu einem Lernmaterial vor einem Schlusstest zu demselben Lernmaterial durchführen (Adesope, Trevisan & Sundararajan, 2017). Im Vergleich zu Lernenden ohne einen solchen Test schneiden erstere besser ab (vgl. auch Wood, 2009). Tests innerhalb von Lernaktivitäten – statt wie vielfach nur am Schluss des Lernprozesses – verbessern das erworbene Wissen von Lernenden stärker (Larsen, Butler & Roediger, 2008). Diese das Lernen begleitenden und nicht promotions- oder selektionsrelevanten (also unbenoteten) Tests werden auch als formative Tests bezeichnet im Gegensatz zu bilanzierenden und damit als Bestandteil von Zeugnissen fungierenden summativen Tests.

Dabei führen mehrere formative Tests zu noch besseren Ergebnissen als nur ein einzelner (Larsen et al., 2008; Roediger & Kapicke, 2006). Ebenfalls zu besseren Ergebnissen führen Testaufgaben, die verlangen, dass die Studierenden darin selbst Antworten konstruieren müssen – anstelle von Ankreuzen einer richtigen Antwort (Multiple-Choice-Test). Noch besser fallen Lernergebnisse aus, wenn auf solche Tests Feedback erfolgt. Sehr gut möglich ist dies beispielsweise mit dem digitalen Lern-Tool kahoot (genauere Informationen unter www.kahoot.com). In deren Classic-Variante können Fragen mit verschiedenen Antworten versehen werden. Den Studierenden können diese Fragen über einen Computer auf einem Screen präsentiert werden. Diese haben dann die Möglichkeit, auf ihrem Handy oder Computer

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07

die korrekte Antwort innerhalb einer vorgegebenen Zeit als Farbfeld zu wählen. Der Wettbewerbscharakter (Punkte gibt es nur für die richtige Antwort – je schneller diese erfolgt, desto mehr Punkte gibt es) aktiviert die Konkurrenz mit den Peers und das spielerische Element. Nach jeder Frage hat die Lehrperson die Möglichkeit für ein ausführliches Feedback zu allen soeben gegebenen Antwortvarianten. Ergänzende Fragen der Studierenden vertiefen das Thema weiter. Es ist auch sehr einfach möglich, denselben Test vor und nach einer Lerneinheit einzusetzen, weil die Ergebnisse stets sofort greifbar sind und die Lehrperson keine Zeit mit Korrigieren verliert. Dieses Tool dürfte das Lernen Studierender in beachtlichem Ausmaß aktivieren.

Bei der Frage, wie nachhaltig durch wiederholte Tests erworbenes Wissen ist (Wood, 2009), interessiert vor allem, ob damit auch der Erwerb von Fertigkeiten unterstützt wird – nicht nur das Speichern von Wissen. Hierzu konnten Kromann, Jensen und Ringsted (2009; vgl. auch Pan & Rickard, 2018) zeigen, dass Lernende, die Fertigkeiten mit Hilfe von formativen Tests erworben hatten, in einem Follow-up-Test zwei Wochen später besser abschnitten, womit diese Art des Lernens auch nachhaltig ist. Besonders interessant ist der Vergleich mit dem Anfertigen von Notizen (Rummer, Schweppe, Gerst et al., 2017). Denn leider wird das wiederholte Testen zu oft mit wiederholtem Lesen verglichen. Dies, obwohl bekannt ist, dass dieses weniger wirksam ist als das Anfertigen von Notizen. Letzteres ist in sehr vielen Studienfächern sehr bedeutsam, auch in der Ausbildung von Lehrpersonen. Rummer et al. (2017) haben deshalb verglichen, ob Lernende Wissen besser behalten, wenn sie Notizen anfertigen, als wenn sie wiederholt lesen und wiederholt Tests durchführen. Unmittelbar nach der Intervention zeigte das Anfertigen von Notizen die stärksten Behaltensleistungen, eine Woche später waren die Gruppen mit angefertigten Notizen und mit wiederholtem Testen signifikant besser als die Gruppe mit wiederholtem Lesen, zwei Wochen später jedoch war nur noch die wiederholt getestete Gruppe besser als die anderen beiden. Dies ist gerade auch für angehende Lehrpersonen von Bedeutung, wird in deren Veranstaltungen doch viel Wert auf das Anfertigen von Notizen gelegt. Die Präferenz von Ausbildenden für das Anfertigen von Notizen im Vergleich zum wiederholten Testen sollte deshalb vermehrt Gegenstand von Weiterbildung werden.

Adesope et al. (2017) haben zur Frage nach der Wirkung von Übungstests eine Meta-Analyse durchgeführt und konnten zeigen, dass

Übungstests allen anderen Bedingungen überlegen waren. Die Effektgrößen fielen mehrheitlich moderat aus. Dabei war wiederholtes Testen auch wirksamer als die Wiederholung des erworbenen Inhalts.

Eine der wichtigsten Fragen für das Lehramtsstudium ist die zur Übertragbarbarkeit von erworbenem Wissen. Hierzu haben Pan und Rickard (2018) eine Meta-Analyse zu Transfereffekten durch Übungstests durchgeführt, die sich vor allem auf den Erwerb von Kompetenzen für Gesundheitsberufe bezog. Sie umfasste 192 Effektstärken aus 189 Studien aus mehr als 40 Jahren mit insgesamt mehr als 10.000 Teilnehmenden. Im Vergleich zu verschiedenen Kontrollbedingungen ohne Übungstests fand sich eine beachtliche Effektstärke von d=0.4. Besonders ertragreich waren Übungstest-Formate mit Anwendungs- und Schlussfolgerungsfragen bei Problemen, welche den Einbezug von diagnostischem Wissen erforderten. Die schwächsten Effekte fanden sich für Übungstests mit Stimulus-Response-Items, mit während dem vorangehenden Studium zwar präsentierten jedoch nicht getesteten Inhalten, und mit Problemen, die sich auf ausgearbeitete Fallbeispiele bezogen. Zudem zeigte sich eine hohe Übereinstimmung mit anfänglichen Testleistungen (Ergebnis im ersten Übungstest) sowie gute Ergebnisse für elaborierte Abruf-Expertise. Insgesamt zeigt sich, dass Übungstests das Lernen substanziell verbessern.

Schwieren, Barenberg und Dutke (2017) konnten den Testing-Effekt, wonach Übungstests zu besseren Lernerfolgen führen, ebenfalls für ertragreiches Lernen in Psychologieklassen nachweisen, beim Erwerb von Kompetenzen also, die auch für den Lehrberuf von hoher Bedeutung sind. Sie konnten diesen Testing-Effekt als robusten empirischen Befund bestätigen. Danach erleichtern Übungstests den späteren Abruf im Langzeitgedächtnis. Im Zentrum jüngerer Studien stand dabei die Übung des Abrufs aus dem Langzeitgedächtnis. Es wurden 19 Studien einbezogen, welche diesen Effekt beim Lehren und Lernen von Psychologie prüften. Das Ergebnis bestand aus total 72 Effektstärken, mit einem signifikanten und sehr beachtlichen Over-All-Effekt von der Stärke d = 0.56 für Übungstests in Psychologieklassen.

Bilanzierend kann festgehalten werden, dass der Einsatz von formativen Tests den Erwerb unterschiedlicher Kompetenzen nachhaltig unterstützt und gegenüber vielen Alternativen stärkere Wirkungen erzielt.

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07

Literatur

Alderson, J. C. & Wall, D. (1993). Does Washback Exist? *Applied Linguistics, 14,* 115-129. http://dx.doi.org/10.1093/applin/14.2.115

Adesope, O. O., Trevisan, S. A. & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659-701.

Bastian, J. & Combe, A. (2007). Der Lehrerberuf zwischen öffentlichen Angriffen und gesellschaftlicher Anerkennung. Alltagsbeobachtungen – professionstheoretische Erklärungen – Perspektiven der Schulentwicklung. In N. Ricken (Hrsg.), *Über die Verachtung in der Pädago*gik (S. 235-247). Wiesbaden: VS Verlag für Sozialwissenschaften.

Cheng, L., Curtis, A. (2004). Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In L. Cheng, Y. Watanabe & A. Curtis (Hrsg.), *Washback in Language Testing: Research Contexts and Methods* (pp. 3-18). Mahwah/New Jersey u. a.: Erlbaum Associates.

Helmke, A. (2007). Guter Unterricht – nur ein Angebot? *Friedrich Jahresheft 2007*, 62-63.

Kromann, C., Jensen, M. & Ringsted, C. (2009). The effect of testing on skills learning. *Med Educ*, *43*, 21-7.

Larsen, D. P., Butler, A. C. & Roediger, H. L. (2008). Test-enhanced learning in medical education. *Med Educ*, *42*, 959-66.

Mc Ewen, N. (1995). Educational accountability in Alberta. *Canadian Journal of Education*, *20*, 27-44.

Pan, S. & Rickard, T. C. (2018). Transfer of Test-Enhanced Learning: Meta-Analytic Review and Synthesis. *Psychological Bulletin*, *144*(7), 710-756.

Reid, P. F. (2011). Pre-service elementary teachers' mathematical beliefs and attitudes about high-stakes testing. *Dissertation. Humanities and Social Sciences, Vol 71(10 A).*

Ro, J. (2019). Learning to Teach in the Era of Test-Based Accountability: A Review of Research. *Professional Development in Education*, *45*(1), 87-101.

Roediger, H. L. & Karpicke, J. D. (2006). The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci*, *132*, 354-84.

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.

Schwieren, J., Barenberg, J. & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective. *Psychology Learning and Teaching*, *16*(2), 179-196.

Smith, W. & Holloway, J. (2020). School testing culture and teacher satisfaction. *Educational Assessment, Evaluation & Accountability*, *32*(4), 461-479.

Wood, T. (2009). Assessment not only drives learning, it may also help learning. *Medical Education*, *43*, 5-6.

Bernhard Hauser, Prof. Dr. phil.
an der Pädagogischen Hochschule St. Gallen.
Arbeitsschwerpunkte:
Lehren und Lernen,
Lernwirksamkeit von Spiel,
Bildungsforschung bei 3- bis 10-Jährigen

bernhard.hauser@phsg.ch

jlb no. 1/2022 https://doi.org/10.35468/jlb-01-2022-07