Robin Woll, Matthias Birkenstock, Daniel Mohr, Pascal Berrang, Tino Steffens, Jörn Loviscach

Hundert Jahre Quizze – und nichts dazugelernt? (Visionen & Konzepte)

Zusammenfassung

In Online-Kursen feiern kurze, automatisierte Lernstandskontrollen („Quizze“) in Form von Multiple-Choice-Tests und Freitexteingaben eine Renaissance. Allerdings werden automatisierte Tests oft zu naiv eingesetzt und bleiben damit hinter ihren Möglichkeiten zurück. Dieser Beitrag ist ein Plädoyer für intelligentere Formen. Er benennt in der Praxis zu wenig beachtete Arbeiten zum klassischen Einsatz von automatisierten Tests als formative Assessments, zeigt aber vor allem neuere Arbeiten und neue Ideen auf, wie automatisierte Tests nicht nur Lernstandskontrollen sein können, sondern direkte Lerneffekte besitzen. Außerdem diskutiert er die schlanke Produktion von Quizzen.

1 Einleitung

Zwar nicht die Selbsttests, aber zumindest die schnell auswertbaren Tests haben eine mindestens hundertjährige Geschichte, wie der „Kansas Silent Reading Test“ von Frederic Kelly (1916) belegt. Fast ebenso lang dauert die Diskussion darüber an: Viele Anwendungen von Multiple-Choice-Tests – in den meisten Fällen eigentlich *Single*-Choice-Tests – und kurzen Freitexteingaben werden zu Recht als geistlos kritisiert. Andererseits sind dies schon allein aus technischer Sicht verlockende Formate, weil sie sich einfach und robust automatisiert prüfen lassen.

Die Komplexität der so prüfbaren Aufgaben ist naturgemäß beschränkt: Wenn der Test zum Beispiel bloß das zahlenmäßige Endergebnis einer aufwändigeren physikalischen Aufgabe prüft, ist der diagnostische Wert gering, denn auf dem Weg dorthin können zu viele verschiedene Fehler passiert sein; meist gibt es gar mehrere mögliche Modellannahmen mit vielleicht sogar verschiedenen Ergebnissen. Hier sorgt die automatische Rückmeldung „Falsch!“ allenfalls für Irritationen. Erst recht sind mehrschrittige Argumentationen, diskutierbare Annahmen und Ergebnisse kaum und gestalterische Aufgaben gar nicht mit automatisierten Selbsttests prüfbar.

Oft wird als Vorteil der Selbsttests deren Objektivität hervorgehoben: Die Bewertung bleibt gleich, unabhängig von der Tagesform der/des Prüfenden und

unabhängig von einem Anker-Effekt durch die davor bewerteten Leistungen. Allerdings misst ein naiv konstruierter Selbsttest im Zweifelsfall nicht das, was eigentlich von Interesse ist. Eine korrekte Antwort kann auch bedeuten, dass es versteckte Hinweise in der Auswahl der Antworten gibt oder dass die Antwort auswendig gelernt wurde, vielleicht sogar ungewollt, zum Beispiel durch Üben mit Probetests oder mit Gedächtnisprotokollen vorheriger Teilnehmer. Reliabilität ist nicht Validität.

Trotz aller dieser Einschränkungen sind automatisierte Selbsttests nicht nur aus technischer Sicht verlockend: Die schnelle Rückmeldung erlaubt, Probleme frühestmöglich zu erkennen (und hoffentlich auch zu beheben). Sie korrigiert eine Unterschätzung oder Überschätzung der eigenen Leistung und stärkt damit das Selbstvertrauen oder nährt nötige Zweifel. Geeignet aufgebaute Selbsttests können durchaus auch Verständnis prüfen, siehe Abschnitt 2 dieses Beitrags.

Selbsttests dienen aber nicht nur der Rückmeldung, sondern können auch zum Aufpassen oder Mitdenken anhalten. Man kann sie als Gamification verstehen, so wie in einem Spiel beständig Hindernisse zu überwinden sind. Sie können direkt dem Lernen dienen – mehr als das erneute Lesen eines Texts – und Denkweisen einüben, siehe den Abschnitt 3 dieses Beitrags. Solche Anwendungen weisen über die üblichen behavioristischen Einsätze von Selbsttests hinaus.

2 Formatives Assessment

Zur „professionellen“ Konstruktion von Selbsttests gibt es viele Ratgeber (siehe die Studie in Haladyna et al., 2002), die sich meist der Abwehr von „Test Wiseness“ (der Fähigkeit, die richtige Antwort zu raten) und der Verständlichkeit widmen, aber oft zentrale Punkte auslassen: 1. Validität: Testet der Test, was er testen soll? (Und welches eine Konzept soll das sein?) 2. Diagnostische Tiefe: Erlaubt das Testergebnis Rückschlüsse auf Probleme? Zum Beispiel sollten Rechenaufgaben so gestellt sein, dass sich aus dem Ergebnis die Art des Fehlers schließen lässt. Schon deshalb verbietet sich die Antwortoption „none of the above“.

Weil die korrekten Antworten von Multiple-Choice-Tests bei manueller Platzierung seltener eine Randposition haben (Attali & Bar-Hillel, 2003), empfiehlt sich eine per Computer ausgewürfelte Platzierung. Das Format „Discrete-Option Multiple-Choice“ geht noch einen Schritt weiter als die per Computer ausgewürfelte Platzierung von Antworten, indem es die Antwortmöglichkeiten eine nach der anderen abfragt, so dass nie die gesamte Liste zu sehen ist (Foster & Miller, 2009).

Wenn angesagt wird, dass nicht mehr nur eine einzige Antwort pro Frage richtig sein muss, kann die Quote richtiger Antworten drastisch fallen (Schulze et al., 2005). In der Mathematik können bei freier, algebraisch geprüfter Formeleingabe Lösungen auf unendlich viele Weisen geschrieben werden ($2 = 6/3 = \sqrt{4} = \ldots$) und sogar unendlich viele verschiedene Lösungen korrekt sein: „Geben Sie die Gleichung einer Gerade an, welche die x-Achse nicht schneidet."

„Ordered Multiple-Choice Tests" prüfen die Verständnistiefe, indem ihre Antwortoptionen jeweils verschiedenen Verständnisstufen entsprechen (Hadenfeld & Neuman, 2012). Freitexteingaben sind dem tieferen Lernen zuträglicher als Multiple-Choice-Aufgaben (Simkin & Kuechler, 2005). Ein Test auf tieferes Verstehen sind „Two-Tier Multiple-Choice Tests", in denen eine Aussage als wahr oder falsch zu bewerten und aus einer Liste von Begründungen eine auszuwählen ist (Treagust, 2006). Eine interdisziplinäre Verbindung: Tests zum Nachweis tiefer Verarbeitung sind auch in der Künstlichen Intelligenz ein Thema (Levesque, 2013).

3 Mehr als Prüfungen

Selbsttests nur zur Überprüfung einzusetzen, heißt, ihr Potenzial zu verschenken: Das Nachdenken über die Frage verursacht einen Behaltenseffekt, der über den des abermaligen Lesens eines Texts hinausgehen kann und obendrein das Selbstvertrauen stärkt (Agarwal et al., 2012). Rückmeldungen – am besten sogar verzögert – über die richtige Antwort erhöhen diesen Effekt und lindern, dass bei einem Multiple-Choice-Test falsche Antworten erscheinen, die später erinnert werden könnten (Butler & Roediger, 2008). Multiple-Choice-Tests können sogar anders als Freitextantworten helfen, falsche Antworten als solche zu erinnern (Little et al., 2012).

Längere Prozeduren, wie man sie etwa mit „worked examples with fading" (Moreno et al., 2006) einübt, lassen sich in Reihen von Quizzen übersetzen (Jeuring et al., 2011; Schypula et al., 2013). Die Einsatzmöglichkeiten von Selbsttests reichen vom Schaffen von Aufmerksamkeit bis hin zu höheren Stufen der Bloom-Taxonomie (Krathwohl, 2002). Schon einfache Quizze helfen, bei Online-Vorlesungen weniger mit den Gedanken abzuschweifen (Szpunar et al., 2013). Auf der anderen Seite kann das Testen auch den Transfer unterstützen (Carpenter, 2012).

Ohne automatische Rückmeldung müssen Leitfragen, Fragen nach der eigenen Einschätzung, Fragen nach Beispielen und ähnliche Fragen auskommen (siehe z.B. Williams, 2013). Man kann sie vielleicht mit Aufforderungen zu Selbsterklärungen vergleichen. Diese sind als lernförderlich bekannt (Roy & Chi, 2005).

Gerade, wenn die Teilnahme an Selbsttests optional ist, erweist sich der dafür verlangte Denkaufwand als Zwickmühle: Die Tests dürfen nicht zu schwierig sein, um die Lerner nicht abzuschrecken; gleichzeitig müssen sie schwierig genug sein, um einen Lerneffekt zu haben, denn gerade der Erinnerungsaufwand fördert das längerfristige Lernen (Bjork & Bjork, 2011). Um diese Situation zu entschärfen, kann man über motivierende Texte à la „Übung macht den Meister!“ nachdenken (Williams et al., 2013), über Adaptivität, Gruppendynamik und Gamification. Letztere kann sich in äußerlichen Spielelementen wie Abzeichen, Punkteständen, Bestenlisten und „Levels“ zeigen (mit zweifelhaftem Nutzen, siehe Ferrara, 2013), aber auch in einem durchdachten Aufbau mit austariertem Schwierigkeitsniveau.

Die Autoren betreiben seit September 2013 die Plattform Capira[1], die Quizze grafisch über Videos legt. So mag ein Mathematik-Video kurz vor dem Resultat einer vorgeführten Rechnung stoppen; im Videofenster erscheint an dem Platz des erwarteten Ergebnisses eine Eingabeaufforderung. Dort soll das Ergebnis vorab eingegeben werden, als Selbsttest. Nach der Eingabe läuft das Video weiter und man erfährt die korrekte Antwort, samt Erklärung.

Die leichtfüßige grafische Verbindung mit dem Video ohne unruhiges Hin und Her zwischen Videoansicht und Frageseiten erlaubt eine höhere zeitliche Dichte von Quizzen. Außerdem lassen sich Quizze an gezielten Stellen in Texten, Bildern und Formeln platzieren, was den Ballast eines Begleittext oft überflüssig macht. Im auf Capira derzeit verfügbaren Mathematik-Brückenkurs wird diese Technik insbesondere eingesetzt, um zum Mitdenken beim Betrachten der Erklärvideos anzuleiten: Die Quizze wiederholen etwa gerade benötigte Grundbegriffe, fragen aber insbesondere nach dem kommenden Schritt der im Fortschritt befindlichen Rechnung.

Eine weitere Anwendung von Quizzen, welche die Autoren derzeit erforschen, ist das Einüben von Heuristiken in der Mathematik. Quizze stellen hier Fragen (mit automatisch geprüften Antworten), die sich Lernende später selbst stellen sollen, zum Beispiel: In welcher Größenordnung liegt das Ergebnis? Was wird der letzte Schritt einer Rechnung sein, mit der man das gewünschte Resultat bestimmt?

4 Produktion von Quizzen

Während etwa Screencast-Vorlesungsmitschnitte heute schnell und einfach machbar sind, erweist sich die Produktion von Selbsttests immer noch als aufwändig. Offensichtlich, um sehr viele Anwendungsfälle abzubilden, geraten die Autorenansichten von Selbsttests in gängigen Learning-Management-Systemen,

1 www.capira.de

aber auch in LearningApps.org[2] und in dem in gewissem Rahmen adaptiven Oppia[3] komplex. Im Unterschied zu den beiden letzteren verlangen die klassischen Learning-Management-Systeme obendrein relativ viele Mausklicks auf Knöpfe wie „Absenden“ und „Speichern“ und verteilen die Eingaben auf viele Bildschirmseiten.

Die Autoren haben deshalb Software entwickelt, die es gestattet, Quizze „live“ mit einem Erklärvideo aufzuzeichnen. Eingabefelder usw. werden direkt am Tablet gezeichnet, im Zweifelsfall sogar, ohne die Aufnahme zu stoppen. Eine über richtig/falsch hinausgehende Rückmeldung zu dem Quiz ist oft verzichtbar, weil direkt danach im Video die Lösung ausgeführt werden kann. Neben dem Eingabeaufwand steht der Aufwand, der in die Didaktik von Quizzen gesteckt werden sollte – aber selten gesteckt wird. Hier arbeiten die Autoren an Mustervorlagen insbesondere für die erwähnten Aufgaben, die Heuristiken einüben sollen.

5 Fazit

Schon die kurzen Selbsttests bieten ein selten ausgeschöpftes Potenzial zum Lernen. Dies geht weit über das reine Abprüfen von elementarem Wissen und Können hinaus. Umso größer werden die Möglichkeiten mit den hier nicht besprochenen längeren, ebenfalls automatisiert prüfbaren Aufgaben, etwa zur Entwicklung elektronischer Schaltungen in einem Simulator oder zur Programmierung, und mit Aufgaben mit Kollaboration und/oder Rückmeldungen unter „Peers“. Diese weitergehenden Möglichkeiten sind allerdings aufwändiger und didaktisch schwerer vorauszuplanen. Andererseits ist gerade die didaktische Engführung durch kurze Selbsttests ein „Spoonfeeding“, das mit vielen Vorstellungen von akademischer Bildung kollidiert, sich aber als Gewohnheit einschleifen kann. Um „Problemlöser(in)“ auf Hochschulniveau werden zu können, darf man sein Studium nicht einsam vorm Rechner sitzend mit automatisiert geprüften Tests absolviert haben.

Literatur

Agarwal, P. K., Bain, P. M. & Chamberlain, R.W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24,* 437–448.

2 http://learningapps.org
3 www.oppia.org

Attali, Y. & Bar-Hillel, M. (2003). Guess where: the position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109–128.

Bjork, E. L. & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M.A. Gernsbacher et al. (Hrsg.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (S. 56–64). New York: Worth Publishers.

Butler, A. C. & Roediger, H. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283.

Ferrara, J. (2013). Games for persuasion. argumentation, procedurality, and the lie of gamification. *Games and Culture, 8*(4), 289–304.

Foster, D. & Miller Jr., H. L. (2009). A new format for multiple-choice testing: Discrete-Option Multiple-Choice. *Psycholgy Science Quarterly, 51*(4), 355–369.

Hadenfeld, J. C. & Neumann, K. (2012). Die Erfassung des Verständnisses von Materie durch Ordered Multiple Choice Aufgaben. *Zeitschrift für Didaktik der Naturwissenschaften, 18*, 317–338.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Jeuring. J., Heeren, B. & Goguadze, G. (2011). *Math-Bridge deliverable D 6.2: diagnostic tools improved and integrated in Math-Bridge.* Abgerufen von http://project.math-bridge.org/downloads/outcomes/deliverables/D6-2.pdf

Kelly, F. J. (1916). The Kansas Silent Reading Tests. *Journal for Educational Psychology, 7*(2), 63–80.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory into Practice, 41*(4), 212–264.

Levesque, H. J. (2013). *On our best behaviour.* Abgerufen von http://www.cs.toronto.edu/~hector/Papers/

Little, J. L., Bjork, E. L., Bjork, R. A. & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges. Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337–1344.

Moreno, R., Reisslein, M. & Delgoda, G.M. (2006). Toward a fundamental understanding of worked example instruction: impact of means-ends practice, backward/forward fading, and adaptivity. In *Frontiers in Education Conference, 36th Annual*, 5–10.

Roy, M. & Chi, M. T. H. (2005). The self-explanation principle. In R. E. Mayer (Hrsg.), *Cambridge Handbook of Multimedia Learning* (S. 271–286). Cambridge: University Press.

Schulze, J., Drolshagen, S., Nürnberger, F., Ochsendorf, F., Schäfer, V. & Brandt, C. (2005). Einfluss des Fragenformates in Multiple-choice-Prüfungen auf die Antwortwahrscheinlichkeit: eine Untersuchung am Beispiel mikrobiologischer Fragen. *GMS Zeitschrift für Medizinischen Ausbildung, 22*(4): Doc218.

Schypula, M., Kurt-Karaoglu, F., Schwinning, N., Striewe, M. & Goedicke, M. (2013). Beobachtungen zur Motivation der Studierenden bei verschiedenen Frageformaten. *DeLFI,* 35–46.

Simkin, M. G. & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*(1), 73–97.

Szpunar, K. K., Khan, N. Y. & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, 110*(16), 6313–6317.

Treagust, D. (2006). Diagnostic assessment in science as a means to improving teaching, learning and retention. *UniServe Science Assessment Symposium*, 1–9.

Williams, J. J. (2013). Improving learning in MOOCs with cognitive science. *AIED 2013 Vol. I, Workshop on Massive Open Online Courses,* 49–54.

Williams. J. J., Paunesku, D., Heley, B. & Sohl-Dickstein, J. (2013). Measurably increasing motivation in MOOCs. *AIED 2013 Vol. I, Workshop on Massive Open Online Courses,* 55.