Mit den Pollen kommt auch PISA wieder. Immer mehr Tests kommen auf den Markt: PISA, BIJU, TIMSS, Iglu, TOSCA, Vera. Immer mehr Menschen werden getestet. Kaum jemand ist noch nicht getestet worden und die Tests folgen in immer schnellerer Folge. Besonders Schüler müssen sich ständig Tests unterziehen, und das weltweit. Testeritis scheint hoch ansteckend zu sein. Kein Wunder: Hinter diesen Tests stecken weltweit aktive Unternehmen, die mit solchen Tests Milliardenumsätze machen. Aber je mehr Tests durchgeführt werden, umso mehr wächst die Unzufriedenheit damit. Was hat man von solchen Tests? Wird dadurch die Schule wirklich besser? Von **Georg Lind**[*]

In den USA boykottieren zunehmend Eltern die Tests. Schüler, Lehrer und ganze Kollegien wehren sich gegen die negativen Folgen von schlechten Testleistungen durch Schwindeln. In Georgia, USA, wurden gerade 170 Lehrer, Schulinspektoren und Schulleiter angeklagt und teilweise zu hohen Gefängnisstrafen verurteilt, weil sie über Jahre hinweg die Noten ihrer Schüler nachbesserten. Die verstorbene Schulrätin, die dabei mitmachte, wurde 2009 noch als beste "Superintendentin der USA" ausgezeichnet. Das Schwindeln hat inzwischen solche Ausmaße angenommen, dass man sich fragen muss, ob man diese Daten wegen der Verzerrung durch Ausfälle überhaupt noch brauchbar sind. Experten fordern einen Teststopp an unseren Schulen, um über Sinn und Unsinn von Schulleistungstests öffentlich zu diskutieren.[1]

Warum sind Tests so in Verruf geraten?
Ich selbst gehörte lange zu den Befürwortern von Tests. Ich habe mit Hilfe von Tests meinen Hochschulunterricht so stark verbessern können, dass der Lernzuwachs der Studierenden vervielfacht wurde.[2] Natürlich haben die Tests selbst keine Wirkung gehabt, sondern die neuen Methoden meines Unterrichts. Aber ohne Tests vor und nach den Unterrichtseinheiten hätte ich nicht feststellen können, ob die neuen Methoden wirklich zu einem Lernzuwachs bei meinen Studenten führten. Wenn sie vernünftig und verantwortungsbewusst eingesetzt werden, sind Tests ein wichtiges Mittel, um den Unterricht und um die Bildungspolitik zu verbessern, so dass alle unsere Kinder eine bessere Ausbildung bekommen. Diese nützliche Wirkung können Tests aber nur haben, wenn sie bestimmte Bedingungen erfüllen, von denen die meisten nicht mehr, sondern eher weniger Kosten verursachen:[3]

Erstens: Tests müssen so valide wie möglich sein. Das heißt, sie müssen wirklich die Lernziele abbilden, um die es im Unterricht geht. Ich schätze, dass 99% der gängigen Tests (PISA, Vera etc.) diese Forderung nicht erfüllen, weil sie mit den falschen Zielen gemacht werden und auf falschen statistischen Modellen beruhen. Diese Modelle optimieren Genauigkeit auf Kosten von Validität, um möglichst viele Unterschiede vorweisen zu

können, auch wenn sie nicht so unbedeutend sind. Bei der Messung von höheren kognitiven Fähigkeiten und bei der Diagnose von Lerndefiziten versagen diese Modelle und die darauf gründenden Tests völlig. Statt dessen sollte die Validität der Tests höchste Priorität erhalten, was man nur dadurch erreichen kann, dass man a) die Erstellung von Testaufgaben von Fachexperten für das jeweilige Testgebiet, Lehrern und Lernpsychologen machen lässt, und b) die Tests so anlegt, dass sie die angezielte Fähigkeit auch wirklich messen, und nicht andere Dinge wie Raten, Testschlauheit oder Angstresistenz.

Zweitens: Das Design der Datenerhebung muss so gewählt werden, dass Vermutungen über zentrale Wirkfaktoren und bekannte Randbedingungen für den Lernerfolg möglichst eindeutig geprüft werden können. Reine "deskriptive" Studien gibt es nicht! Jeder Auswahl von Testaufgaben und jedem Erhebungsdesign liegen Entscheidungen zugrunde, die sich im günstigen Fall an begründeten Erkenntnisinteressen orientieren, öfter leider an rein merkantilen Interessen der Testindustrie: Wie kann ich mit möglichst wenig Aufwand die größtmögliche öffentliche Aufmerksamkeit erregen und mir dadurch weitere Aufträge durch gleichgepolte Politiker sichern? Mit den bisherigen Massenerhebungen an Alterskohorten lassen sich überhaupt keine wichtigen Fragen der Bildungspolitik und des Unterrichts klären. Sie öffnen nur der wilden Spekulation Tür und Tor.

Schon durch relativ einfache Verbesserungen des Erhebungsdesigns ließen sich wichtige Fragen beantworten und Anhaltspunkte für Verbesserungen gewinnen. Schon durch den Vergleich von zwei Klassenstufen, die ein oder zwei Jahre auseinander liegen, kann festgestellt werden, welchen Lernfortschritt der Unterricht bewirkt. Wenn man, was für sinnvolle Vergleich absolut wichtig ist, in den Vergleichsgruppen dieselben Testaufgaben einsetzten würde, könnte man nebenbei noch feststellen, wie nachhaltig die Lerngewinne sind (ob es wirklich stimmt, dass das erworbene Wissen nur bis zur nächsten Klassenarbeit hält). Wenn man dazu noch erheben würde, welche Lernmittel zur Verfügung standen, wie das Ausbildungsniveau der Lehrkräfte ist, aus welchen ökonomischen Verhältnissen die Schüler stammen und auf welche Ressourcen sie zuhause zurückgreifen können, usw. ergäben sich schon viele Anhaltspunkte für pädagogische und bildungspolitische Verbesserungen (siehe unten).

Solche Erhebungen können kostenneutral gemacht werden, da man für gezielte Fragestellungen meist keine so großen Untersuchungsgruppen benötigt wie für Forschung, der nur daran liegt, winzige Unterschiede aufzublähen. Den Kindern ist mehr geholfen, wenn man sich in solchen Erhebungen auf Lernbedingungen konzentriert, die nicht bloß statistisch signifikant, sondern auch "wirksam" sind.

Auswertung und Nutzen

Bei der Auswertung der Erhebungen müssen Effektstärke und Effizienz von Lernbedingungen im Mittelpunkt stehen und nicht Rangplätze, die Unterschiede aufblähen, und statistische "Signifikanz", die mit Bedeutsamkeit nichts zu tun hat, sondern nur mit der Frage, ob die Untersuchungsgruppe groß genug war, um einen noch so winzigen Unterschied aufzuzeigen. Für die Nützlichkeit von Lernleistungsstudien sind eine Reihe weiterer Bedingungen von Bedeutung:

Sachorientierung und Anonymität. Diese Studien müssen auf die Bewertung von Bedingungen des Lernens und von Maßnahmen begrenzt werden, und dürfen nicht dazu missbraucht werden, um Schüler, Lehrer, Schulen und Ländern zu bewerten. Für diese Zwecken müssen die Daten vollständig anonymisiert werden. Dann gäbe es für niemanden einen Grund zu schwindeln. Diese Anonymität schließt nicht aus, dass Daten an individuelle Schulen, Lehrer und sogar Schüler zurück gemeldet werden, wenn sicher gestellt ist, dass personenbezogene Informationen nicht für Dritte zugänglich sind. "Personen" sind in diesem Sinne auch Schulen und Länder, da sie Personenkollektive repräsentieren. Für die Verbesserung des Unterrichts und der Schulen sind solche Informationen auch nicht relevant.

Effektivität
Nach einer allgemein anerkannten Konvention gelten nur solche Faktoren als "wirksam", die mindestens eine Effektstärke von $r > 0.30$ aufweisen. Dieses grobe Kriterium muss im Einzelfall noch durch andere Kriterien ergänzt werden, wie z.B. a) durch das Kriterium der Inklusivität, dass heißt, dass durch eine Lernbedingungen kein Schüler im Lernen behindert werden darf, b) durch das Kriterium der absoluten Effektstärke (z.B. wie viele Schüler schaffen das Minimalziel einer Lerneinheit bzw. wie viele verfehlen es und müssen speziell gefördert werden), und c) durch das Kriterium der Effizienz, das heißt, welche Faktoren können mit dem geringsten Aufwand geändert werden und die größten Effekte bewirken?

Lernfähigkeit
Diese Studien müssen lernfähig sein. Es wäre nicht ungewöhnlich, wenn man nach einer Untersuchung feststellt, dass einige Faktoren völlig unbedeutend sind, oder dass weitere Faktoren zu untersuchen sind. Die Zahl der potentiellen Faktoren ist prinzipiell unbegrenzt, so dass man nie ALLE Faktoren untersuchen kann. Man wird – um Kosten zu begrenzen und um Schüler nicht zu überfordern – in einem Testzyklus immer nur wenige Faktoren berücksichtigen können, und zwar jene, von denen man aufgrund des gegenwärtigen Wissensstands vermuten kann, dass sie die wirksamsten sind. Aber bei jedem neuen Testzyklus können die Faktoren, die man untersucht, ausgetauscht werden, um neue Erkenntnisse zu erhalten.

Anwendungsorientierung
Diese Studien müssen sich auf jene Wirkfaktoren konzentrieren, die von der Bildungspolitik und der Pädagogik auch gestaltet werden können:

- **Effizienz des Unterrichts:** Wie viel Lernzuwachs zeigen die Schüler in einem Jahr? Wichtig ist, dass man in den zu vergleichenden Klassenstufen dieselben Testaufgaben einsetzt! Nach vielen Jahren positiver Erfahrungen damit, kann ich viele Bedenken nicht teilen.

- **Inklusivität:** Wie viele Schüler erreichen die Mindestziele des Unterrichts, wie viele werden vom Unterricht nicht erreicht?

- **Lernmittel:** Wie wirken sich bestimmte Lehrbücher auf den Lernzuwachs aus? Wie fördern oder behindern bestimmte Lernmittel (Verfügbarkeit von Lehrbüchern, Whiteboard, Computer, audio-visuelle Medien) den Lernzuwachs?

- **Lehrkompetenz:** Wie stark wirkt sich der Ausbildungsstand der Lehrkräfte auf den Lerngewinn der Schüler aus? Beim Ausbildungsstand wären mehrere Faktoren zu erfassen: Alter/Erfahrung der Lehrkräfte, Umfang der Fortbildung der Lehrkräfte in Person-Stunden, Altersmischung der Lehrkräfte (was große Auswirkungen auf die Innovations- und Lernfähigkeit eines Kollegiums haben kann), Gesundheitsstand, Bezahlung usw.

- Zudem sollten **Randbedingungen** erhoben werden, auch wenn man sie nur schwer beeinflussen kann, um die obigen Ergebnisse bessere einschätzen zu können und um Änderungen in anderen Politikbereichen anzuregen.

- **Ressourcen der Schüler:** Wie fördern oder behindern der ökonomische Hintergrund, die Beherrschung der Landessprache und andere soziale Bedingungen den Lernzuwachs?

Rückmeldung und Folgen
Ein Test, dessen Befunde nicht direkt Verbesserungen nahe legen, ist ein schlechter Test. Die ehemalige Bildungsministerin von Kolumbien, Velez, sagte mir, sie habe die Sprachkompetenz aller Englischlehrer testen lassen, aber nicht, um die schlechten, wie sonst üblich, an den Pranger zu stellen, sondern um ihnen einen mehrwöchigen Fortbildungskurs zu spendieren. Solche relevanten Testdaten würden, im Unterschied zu PISA & Co bei den Lehrpersonen, Schulleitern und Schuladministratoren Nachdenken und

Diskussionen über Verbesserungen anstoßen. Wenn z.B. eine Lehrperson feststellt, dass ihre Klassen in Mathematik deutlich unter dem Lerngewinn bleiben, die andere erzielen, könnte sie sich fragen, ob sie das richtige Lehrbuch benutzt, oder ob sie vielleicht mal wieder an einer Fortbildungsmaßnahme teilnehmen sollte. Die Freigabe solcher Daten muss natürlich so begrenzt werden, dass die Beteiligten nur eigene Daten bekommen. Dieser Service wäre ein Ausgleich dafür, dass die teilnehmenden Schulen, Lehrpersonen und Schüler einen Teil ihrer Zeit für die Testdurchführung opfern. Er würde auch die Beteiligten motivieren, die Tests ernst zu nehmen und gewissenhaft zu bearbeiten.

Ausfälle vermeiden
Nur Tests, an denen nahezu alle teilnehmen und die von Allen motiviert ausgefüllt werden, haben einen Wert. Schon geringe Ausfälle können die Ergebnisse so verzerren, dass sie wertlos sind, besonders wenn man sich auf das Auffinden von winzigen Unterschieden konzentriert. Aber auch Fragen nach substantiellen Effekten sind betroffen, wenn die Ausfälle – in Folge von individuellen Verweigerungen oder organisierten Boykottmaßnahmen – ein größeres Ausmaß annehmen. Ausfälle verzerren die Ergebnisse oft so stark, dass sie zu falschen Schlussfolgerungen und Maßnahmen führen. Wenn nur zehn Prozent der Schüler die Tests boykottieren, sollte man den Ergebnissen nicht mehr trauen.

Bisherige Praxis nicht nur Geldverschwendung, sondern schädlich
Die Bildungspolitik muss sich unbedingt kompetent machen, bevor sie weitere Schulleistungstests in Auftrag gibt. Die bisherige Praxis ist nicht bloß Geldverschwendung. Sie fügt der Bildung an unseren Schulen auch großen Schaden zu. Zum Glück, muss man sagen, besteht ein gewisses “Handlungsdefizit”. Der Schaden wäre vermutlich noch größer, wenn alle Empfehlungen der OECD und der PISA-Konsorten eins zu eins umgesetzt würden.

Schulleistungsstudien müssen transparent für die Öffentlichkeit sein, damit sie kritisch analysiert und aufgrund von Diskussionen valider gemacht werden können: Messen die Testaufgaben wirklich das, was sie messen sollen? Dienen Tests wirklich der Verbesserung des Unterrichts und der Bildungspolitik? Oder werden dadurch nur einige multinationale Testfirmen reicher, die immer mehr Macht über den Unterricht und den Schulbetrieb in den USA und anderen Ländern gewinnen, und zunehmend auch bei uns?

[«*] **Prof. em. Dr. Georg Lind** – Konstanz

[«1] Jahnke, T. & Meyerhöfer, W., Hg. (2006). Pisa & Co. Kritik eines Programms.

Hildesheim: Franzbecker Verlag.
Hopmann, S. T., Brinek, G., & Retzl, M., Hg. (2007). PISA zufolge PISA. PISA according to PISA. Hält PISA, was es verspricht? Wien: Lit-Verlag.
Lind, G. (2009). Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In: T. Bohl & H. Kiper., Hg.: Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren, S. 61 – 79. Bad Heilbrunn: Julius Klinkhardt.

[«2] Lind, G. (2015). Favorable learning environments for moral competence development [PDF – 1.3 MB] – A multiple intervention study with nearly 3.000 students in a higher education context. International Journal of University Teaching and Faculty Development, 4, 4.

[«3] Lind, G. (2004). Jenseits von PISA — Für eine neue Evaluationskultur. In: Institut für Schulentwicklung PH Schwäbisch Gmünd, Hg., Standards, Evaluation und neue Methoden. Reaktionen auf die PISA-Studie, S. 1 – 7. Baltmannsweiler: Schneider Verlag Hohengehren.
Lind, G. (2011). Verbesserung des Unterrichts durch Selbstevaluation. Ein Plädoyer für unverzerrte Evidenz. In: J. Bellmann & T. Müller, Hg.: Wissen, was wirkt. Kritik evidenzbasierter Pädagogik, S. 173 – 195. Wiesbaden: VS-Verlag für Sozialwissenschaften. ISBN 978-3-531-17688-8
Brügelmann, H. (2015): Vermessene Pädagogik – standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co. Beltz: Weinheim/ Basel (im Druck).