

Lind, G. (2019). Lernstandserhebungen und Diagnosen JA, aber WIE? Eine Kritik der herrschenden Testverfahren und der 'Evidenz'-Basierung. In: Lehren & Lernen, S. 17-19, 5-2019.

Georg Lind

## Lernstandserhebungen und Diagnosen JA, aber WIE?

### Eine Kritik der herrschenden Testverfahren und der „Evidenz“-Basierung

(Anonyme) Daten, die für die Verbesserung von Lernen, Unterricht und Bildungspolitik benötigt werden, müssen nur richtig erhoben und interpretiert werden. Personenbezogene Daten, die der Benotung von Menschen dienen, müssen jedoch höhere Anforderungen erfüllen, da sie ungewollte Nebenwirkungen haben und in Lernprozesse massiv eingreifen.

► Stichwörter: Lernstandserhebungen, Diagnoseverfahren, Bildungspolitik, Testverfahren, Evidenzbasierung

Die (anonyme) Überprüfung von Lernergebnissen und von Ergebnissen schulischen Unterrichts sowie Maßnahmen von Bildungs- und Schulpolitik ist unverzichtbar, da sie den Anbietern von Bildung (Lehrpersonen, Schulverwaltung, Schulpolitik) Rückmeldung gibt, ob das Richtige in ausreichendem Maß getan wurde und wird.

#### Wie sollte vorgegangen werden? Überprüfen JA – Benotung NEIN

■ Wenn **Schüler/innen** wissen möchten, ob sie ein bestimmtes Lernpensum erfolgreich abgeschlossen ha-

ben, helfen vor allem **freiwillige Probeerarbeiten**, die sich an Lehrbuchkapitel oder Unterrichtseinheiten anschließen.

■ Wenn **Lehrkräfte** wissen möchten, ob sie einen Stoff oder eine Kompetenz erfolgreich an ihre Schüler/innen vermitteln konnten und ob einige Schüler/innen besondere Förderung benötigen, leisten vor allem **kurze, nicht benotete Tests** sehr gute Dienste, die sich aus leichten, mittelschweren und schweren Aufgaben vergangener und zukünftiger (!) Klassenarbeiten zusammensetzen.

■ Wenn **Bildungsverwaltungen** wissen möchten, ob irgendwo Hilfe notwendig ist und ob die eigene Bil-

dungspolitik überdacht werden muss, können sie das auch an **kurzen, anonymen Stichproben-Erhebungen** erkennen. Diese können sich z. B. auf Klassenstufen vor wichtigen Weichenstellungen (z. B. 3, 9 und 11) konzentrieren, so dass Maßnahmen nicht zu spät kommen.

- I Wenn **weiterführende** Schulen herausfinden wollen, welche Kenntnisse der Unterricht in übergangskritischen Fächern wie Mathematik und Englisch bei den Schüler/innen voraussetzen kann und wo Ausgleichsmaßnahmen notwendig sind, können sie am Anfang von Klasse 5 bei allen Schülern **nicht benotete Kurztests** durchführen.
- I Wenn man wissen will, warum **manche Schüler/innen** im Unterricht nicht mitkommen, sollten sie auf die Funktionsfähigkeit ihrer Fernsinne (Augen, Gehör), aber auch ihrer Nahsinne (Mund, Finger) hin geprüft werden, so dass man durch **Ausgleichsmaßnahmen** (Brille, Hörgerät) und Training (Mund-Finger-Training) Abhilfe schaffen kann.

Nach allem, was wir wissen, sind solche **nicht benoteten Erhebungen** gut geeignet, um die Lehrkräfte in ihrem Bemühen um guten Unterricht und gute Schulen zu unterstützen und um Schülern maximale Anregung und Gelegenheit zum Lernen zu geben. Wenn unsere Schulen mit mehr Computern und digital-kundigen Lehrkräften ausgestattet werden, können viele dieser Erhebungen online erfolgen und so die Lehrkräfte stark entlasten.<sup>1</sup>

## Benotete Prüfungen

Auch **umfangreiche benotete Prüfungen** haben eine wichtige Funktion. So macht es Sinn, dass Sekundarschüler **am Ende ihrer Schulzeit** an einer gründlichen Prüfung des Wissens und der Fähigkeiten teilnehmen müssen, die für die Bewältigung einer anschließenden Ausbildung und ganz allgemein ihres bürgerlichen Lebens wichtig sind.

Aber hier kommt es auf die **Dosis** an. Sie sollten **so wenig wie möglich** eingesetzt werden, da sie sonst das Lernen eher behindern als fördern. Sie müssen umfangreicher sein als die eben genannten Prüfungsarten, damit sie eine Note begründen können, und sie müssen auch viel höheren Ansprüchen an ihre Objektivität und Validität genügen.

Von zu vielen, benoteten und dazu noch schlecht gemachten Prüfungen geht eine Gefahr für den Erfolg schulischen Lernens aus. Mehr Druck und mehr Prüfungen können zwar helfen, den Schulbetrieb aufrechtzuerhalten, aber bessere Lernleistungen können sie nicht bewirken. Der Ruf nach mehr Druck und mehr Prüfungen sollte als Hinweis für Defizite in der Lehrerbildung angesehen werden.

Bei allen Prüfungen, denen sich Menschen zwangsweise unterziehen müssen und die Folgen für ihren Ruf und ihre

Lebenschancen haben, sind **strenge Qualitätsmaßstäbe** anzulegen. Vor allem ist es enorm wichtig, dass sie genau das prüfen, was sie zu prüfen vorgeben, also „valide“ sind. Dafür ist unverzichtbar, dass der Prüfungsgegenstand (Wissen und Fähigkeiten in diversen Fächern und auch überfachliche Fähigkeiten) so genau und klar definiert und beschrieben werden, dass sie objektiv geprüft werden können, wobei die Form der Prüfung keine Rolle spielt (Test, Klausur, mündliche Prüfung, Projekt usw.), **d. h. dass jeder Beteiligte anhand des Augenscheins und von Daten erkennen kann, ob das Richtige und ob richtig geprüft wird. Diese Fragen lassen sich nicht rein statistisch ohne inhaltlichen Sachverstand beantworten.**

Bei aller Kritik im Detail erfüllen bisherige Abschlussprüfungen diese Kriterien recht gut, jedenfalls besser als viele der in den letzten Jahren aufkommenden „objektiven“ und „reliablen“ Tests. Es stimmt, dass trotz hoher Standardisierung die Einschätzungen der Prüfungsleistungen durch Lehrpersonen nicht perfekt sind. Die Kriterien für die Vergabe von Punkten sind oft sehr subjektiv (vor allem in den Sprachfächern). Die Bewertungen verschiedener Prüfer schwanken oft stark.

## Was ist von „Evidenz“-Basierung zu halten?

Bisher als Alternative vorgeschlagene „Evidenz“-basierte Tests sind im Vergleich dazu noch problematischer, weshalb sie an vielen Orten der Welt bei Eltern, Lehrkräften und Schüler/innen und auch bei Gerichten auf immer größere Kritik stoßen. Problematisch an diesen „Evidenz“-basierten Tests ist u. a.:

- I Die „**psychometrischen**“ Tests beruhen nicht, wie man meinen könnte, auf reiner Evidenz oder auf psychologischen Erkenntnissen, sondern auf **willkürlichen statistischen Modellen**. Oft widersprechen sie diesen sogar: Die Testmacher gehen davon aus, dass die Lösung einer Aufgabe oder die Antwort auf eine Testfrage nur von einem einzigen Faktor abhängen, nämlich von der Fähigkeit, die der Test messen soll. **Aber die Psychologie weiß, dass das Verhalten meist von mehreren Faktoren determiniert sein kann. Das heißt, die Testwerte können verschiedene Dispositionen widerspiegeln und nicht nur die Fähigkeit, die ein Test messen soll.**
- I Diese Annahme (einer einfachen Antwort-Fähigkeits-Beziehung) ist nur durch das **lineare statistische Modell** begründet, für das sich die Testproduzenten entschieden haben, für das es aber sonst **keine pädagogische oder psychologische Begründung** gibt.
- I Die Stabilität (Reliabilität) von Testwerten wird eigenwillig als Eigenschaft des Tests und nicht der Testteilnehmer interpretiert. Testhefte sind immer reliabel (auf Deutsch „unveränderlich“). **Was sich verändert, sind die Antworten der Testteilnehmer, weil ihre Fähigkeiten und Motivationen schwanken können. Es**

gibt keinen überzeugenden Grund, solche Schwankungen dem Test als Messfehler zuzuschreiben.

„Korrelation“ wird als Evidenz für Stabilität angesehen. Statistische Korrelation hängt nicht nur von der Stabilität der Antworten, sondern auch von ihrer Streuung ab. Wenn z. B. alle Testteilnehmer eine Aufgabe richtig lösen bzw. alle dieselbe Antwort geben, ist die Korrelation NULL!

Auch für „internale Konsistenz“, einem alternativen Indikator für die Reliabilität, werden Korrelationen berechnet, und zwar zwischen den Lösungen bzw. Antworten in einem Test. Auch hier werden die Zusammenhänge dogmatisch als ein Merkmal des Tests interpretiert.

Damit verbaut man sich aber die Sicht auf ein ganz wichtiges Merkmal menschlichen Verhaltens, nämlich auf die Struktur der individuellen Persönlichkeit. Ohne die Erfassungen von Verhaltensmustern oder -strukturen können wir Kompetenz nicht richtig verstehen und erfassen. Eine Testtheorie, die individuelle Kompetenzstrukturen als Eigenschaften eines Tests und als „Messfehler“ abtut, macht die Erfassung von Kompetenzen unmöglich.

Diese Kritik an der herrschenden Test-Theorie und die oben genannten besseren Wege wird bereits seit langem

von kompetenter Seite vorgetragen, offenbar aber ohne Wirkung auf die Test-Industrie und ihre Kunden.<sup>2</sup> Es geht nicht um einen akademischen Meinungsstreit, sondern um die Zukunft unserer Schulen.

Wir haben – trotz berechtigter Kritik und trotz Alarmissimus aus der PISA-Ecke – eines der besten Schulsysteme der Welt. Man braucht nur Austauschschüler zu fragen, die das deutsche System aus eigenem Erleben mit anderen vergleichen können. Wir können es durch Tests und Evaluation noch verbessern, wenn wir dabei, wie oben dargelegt, klug und umsichtig vorgehen. Wir können es aber auch zugrunde richten, wenn wir den Statistik-basierten Diagnosen noch mehr Macht über unsere Schulen und Hochschulen einräumen. Sie haben sich bisher als weitgehend wirkungslos erwiesen. Wehe, wenn sie Wirkung haben.

### Anmerkungen

- <sup>1</sup> Weitere Anregungen habe ich auf dieser Webseite zusammengetragen: [http://www.uni-konstanz.de/ag-moral/evaluation/itse\\_home.htm](http://www.uni-konstanz.de/ag-moral/evaluation/itse_home.htm)
- <sup>2</sup> Literatur dazu habe ich hier aufgelistet: <http://www.uni-konstanz.de/ag-moral/evaluation/itse-references.htm>

Prof. i. R. Dr. Georg Lind  
Universität Konstanz  
[Georg.Lind@uni-konstanz.de](mailto:Georg.Lind@uni-konstanz.de)

## Walter Herzog

### Kritik der evidenzbasierten Pädagogik

Das Hauptanliegen evidenzbasierter Pädagogik liegt in der Optimierung pädagogischer Interventionen und Programme durch methodisch streng kontrolliertes empirisches Wissen über deren Wirksamkeit. Evidenzbasierte Pädagogik sollte nicht mit empirischer Bildungsforschung gleichgesetzt werden. Sie stellt auch kein eigenes Paradigma empirischer Erziehungswissenschaft dar, da ihr Fokus in der pädagogischen Praxis liegt. Das Verständnis des Verhältnisses von Wissenschaft und Praxis, das ihr zugrunde liegt, wird in dreifacher Hinsicht kritisiert: hinsichtlich der technologischen Auffassung pädagogischer Praxis, hinsichtlich der suggerierten Theorieabstinentz der pädagogischen Forschung und hinsichtlich der Missachtung der Kommunikation als Medium pädagogischer Wirksamkeit. Der Beitrag schließt mit einer skizzenhaften Reminiszenz an ein alternatives Verständnis von Wissenschaft und Praxis, wie es sich in der Pädagogischen Psychologie findet.

► Stichwörter: **Evidenz, Experiment, Kausalität, Praxis, Wirksamkeit**

### 1. Einleitung

In der Beilage einer schweizerischen Tageszeitung erschien unlängst ein Artikel mit dem Titel „Evidenz-

basierte Personalauswahl – Klarheit statt Handgelenk mal Pi“ (Hell 2011). Darin schreibt der Autor, bei der Auswahl von Personal hätten sich über Jahre und Jahrzehnte unterschiedliche Methoden der Selektion eta-