Pisa-Studie

Falsche Erwartungen an schulische Leistungstests

Gastkommentar
von WALTER HERZOG

Die Ergebnisse der jüngsten Pisa-Studie haben ein geringes Medienecho ausgelöst (NZZ 7. 12. 16). Dafür mitverantwortlich sind die in der Schweiz für Pisa zuständigen Personen der Erziehungsdirektorenkonferenz (EDK) und des Staatssekretariats für Bildung, Forschung und Innovation (SBFI), die eine Stellungnahme verweigerten, da die Qualität der Daten ungenügend sei. Ob es wirklich um die Datenqualität geht oder nicht eher um die im Vergleich zu 2012 unerwartet schlechten Ergebnisse in allen getesteten Bereichen, ist unklar. Das eigentliche Problem liegt aber sowieso anderswo, nämlich bei der Fehleinschätzung der Leistungsfähigkeit von schulischen Leistungstests.

So erwartet die EDK von der Messung der Schülerleistungen im Rahmen des Harmos-Projekts «empirisch gesichertes Wissen hinsichtlich des tatsächlich erreichten Kompetenzniveaus der Schülerinnen und Schüler». Und vom Bildungsmonitoring, das sie gemeinsam lanciert haben, erwarten EDK und SBFI «wissenschaftlich gesicherte Erkenntnisse» über das schweizerische Bildungssystem. Anvisiert wird eine «evidenzbasierte Bildungspolitik», die es erlauben soll, bildungspolitische Entscheidungen auf einer rein zweckrationalen Basis zu treffen. Wenn sich jedoch etwas mit Gewissheit sagen lässt, dann, dass Wissenschaft und Forschung nicht in der Lage sind, die Erwartungen von Politik und Verwaltung einer «sicheren» Grundlage für ihre Entscheidungen einzulösen. Insofern könnten die Ergebnisse von Pisa 2015 von heilsamem Nutzen sein, da sie Anlass bieten, die längst fällige Auseinandersetzung über die Möglichkeiten und Grenzen von Schulleistungstests zu führen.

Die Erwartung, mithilfe von Tests lasse sich sicheres Steuerungswissen für das Bildungssystem gewinnen, ist naiv. Die Sozial- und Erziehungswissenschaften verfügen nicht über Messinstrumente, die auf einem vergleichbar anspruchsvollen Niveau Daten generieren lassen, wie dies bei der Messung physikalischer Grössen der Fall ist. Zudem ist es praktisch nie möglich, das interessierende Phänomen, wie zum Beispiel die Kompetenz eines Schülers, direkt zu messen. Alles, was dem Konstrukteur eines pädagogischen Tests zur Verfügung steht, um zu messen, was er messen will, ist das Verhalten des Testnehmers, d. h. dessen Performanz. Dementsprechend kann die Gültigkeit einer Messung höchst umstritten sein. Denn wie überprüft man, ob ein Test misst, was er zu messen vorgibt, wenn der Messgegenstand nur indirekt zugänglich ist? Da ein Vergleich mit der Realität nicht möglich ist, kann letztlich nur ein weiterer Test etwas über die Qualität der Messung aussagen. Das heisst auch, dass selbst das Messniveau eines pädagogischen Tests nicht eindeutig bestimmt werden kann.

Hinzu kommen weitere Probleme pädagogisch-psychologischer Messungen, wie etwa dasjenige der Reaktivität. Einer Landschaft ist es egal, ob und wie sie vermessen wird, ein Mensch wird sich immer auf irgendeine Weise darauf einstellen, dass er einem Test unterworfen wird. Auch die Vertrautheit mit einem Testformat kann eine wesentliche Rolle spielen.

Solche Einflüsse ausschalten zu wollen, indem man Testformat und Testinhalt unverändert lässt, wäre nicht empfehlenswert. Denn Leistungstests müssen laufend angepasst werden, weil sie sonst ihre ohnehin schon prekäre Messqualität vollends einbüssen würden.

Ein Beispiel kann dies illustrieren: Beim Vergleich der Testergebnisse verschiedener US-Gliedstaaten über mehrere Jahre fiel auf, dass sich die Leistungen der Schülerinnen und Schüler durchwegs verbessert hatten, und zwar so stark, dass sie schliesslich in der Mehrheit der Staaten über dem nationalen Durchschnitt lagen – ein mathematischer Unsinn. Bei näherer Analyse stellte sich heraus, dass die Tests Jahr um Jahr unverändert eingesetzt worden waren, was dazu geführt hatte, dass die Testaufgaben bekanntwurden, die Lehrpersonen ihre Schüler auf die Tests vorbereiteten und die Prüfstichproben mit der Eichstichprobe nicht mehr übereinstimmten. Der vermeintliche Anstieg der Schülerleistungen war ein methodisches Artefakt.

In den Griff bekommen lässt sich der Lake-Wobegon-Effekt (wie er mittlerweile genannt wird) nur, wenn Tests erstens ständig à jour gehalten werden und wenn zweitens Ergebnisse aus mehreren Tests verfügbar sind, um die Qualität eines Tests vergleichend zu beurteilen.

All dies ist nicht als Argument gegen Leistungstests an unseren Schulen gedacht. Jedoch braucht es eine Art nachholende Aufklärung darüber, was schulische Leistungstests leisten können. Sie können sicher nicht leisten, was in den Köpfen einiger Bildungspolitiker und gewisser Vertreter in den Bildungsverwaltungen herumschwirrt, nämlich eine Entpolitisierung bildungspolitischer Entscheidungen durch «wissenschaftlich gesicherte Erkenntnisse».

Walter Herzog ist em. Professor für pädagogische Psychologie an der Universität Bern.