Hat uns eine Testkultur (gerade) noch gefehlt? Möglichkeiten und Grenzen schulischer Leistungstests*

Walter Herzog

Im Unterschied zu den USA und anderen Ländern kennen wir in der Schweiz keine Tradition der Testung von Schülerleistungen. Seit den ersten PISA-Erhebungen gibt es aber Stimmen, die meinen, dass auch wir eine schulische Testkultur brauchen. Externe Leistungstests, wie sie von HarmoS vorgesehen sind, würden den Schulen, so heisst es zum Beispiel bei Jürgen Oelkers (2007), den «Einstieg in die Testkultur» (S. 32) ermöglichen. Dabei suggeriert der Begriff der Kultur, dass es sich um etwas Wertvolles handelt. Die Kultur edelt den Menschen und hebt ihn von den Tieren ab – auch wenn inzwischen Zweifel bestehen, ob die Kultur tatsächlich ein Humanspezifikum darstellt. Darüber will ich aber nicht richten, sondern der Frage nachgehen, ob wir Tests tatsächlich brauchen, der «Einstieg in die Testkultur» also etwas ist, was den Schulen noch gefehlt hat.

«Noch gefehlt hat» dürfen Sie in der Doppeldeutigkeit nehmen, wie der Ausdruck daher kommt: Das hat uns *noch* gefehlt, weshalb wir es schleunigst brauchen, oder: Das hat uns *gerade* noch gefehlt, als ob wir nicht schon genügend unnötige Reformen zu ertragen hätten.

Bevor wir entscheiden können, welches die richtige Deutung ist, müssen wir uns ein Bild von Tests und deren Verwendung machen. Mein Referat ist daher wie folgt aufgebaut: (1) Wozu sollen Tests eingeführt werden? (2) Test ist nicht gleich Test. (3) *High-Stakes Testing* in den USA. (4) Die Situation in der Schweiz.

1. Wozu sollen Tests eingeführt werden?

Die Einführung von Tests an unseren Schulen wird aus verschiedenen Gründen propagiert. Die drei wesentlichen sind m.E. die folgenden:

* Referat anlässlich der Sektionsversammlung des VPOD Zürich Lehrberufe vom 1. Dezember 2012 in Zürich.

• Monitoring des Bildungssystems, inkl. externe Evaluation von Schule und Unterricht
• Objektivierung der Schülerbeurteilung (insbes. der Notengebung)
• Kontrolle und bessere Steuerung des Bildungssystems

1.1 Monitoring

Wie Sie wissen, haben die EDK und das BBT ein schweizerisches Bildungsmonitoring aufgezogen, in dessen Rahmen alle vier Jahre ein umfassender Bericht erarbeitet wird, der über den Stand sämtlicher Bildungsbereiche in der Schweiz Aufschluss gibt. Ein Pilotbericht ist 2006 erschienen. Der erste reguläre Bericht wurde 2010 vorgelegt. 2014 ist der nächste zu erwarten. Erstellt wird er von der Schweizerischen Koordinationsstelle für Bildungsforschung in Aarau.

Im Zentrum der Berichterstattung stehen drei Beurteilungsbereiche, nämlich die Effektivität (Wirksamkeit), die Effizienz (der optimale Einsatz der Ressourcen) und die Equity (Bildungsgerechtigkeit) des schweizerischen Bildungswesens. Dies alles sind Outputkriterien (wenn auch bei der Effizienz der Input mit berücksichtigt wird), d.h. das schweizerische Bildungssystem wird danach beurteilt, was es an Leistungen erbringt. Ein wesentlicher Aspekt der Leistungen sind die Lernleistungen der Schülerinnen und Schüler, welche die Schulen in Form von Noten ausweisen, welche aber im Rahmen des Bildungsmonitorings «objektiv» erfasst werden sollen, d.h. mittels standardisierter Messverfahren. Im Bildungsbericht 2010 finden Sie denn auch keine Noten, aber Ergebnisse von PISA und anderen Schulleistungstests.

Tests werden also für Monitoringzwecke benötigt. Ein Monitor ist ein Überwachungsgerät. Wie in öffentlichen und privaten Räumen Monitore zurzeit in wachsender Zahl installiert werden, reiht sich das Bildungsmonitoring von EDK und BBT in die Tendenz nach mehr Überwachung in unserer Gesellschaft ein. *Big brother is watching you*. Etwas polemisch formuliert, ist es der Überwachungsstaat, der für schulische Leistungstests und die Einführung einer Testkultur an unseren Schulen einsteht.

1.2 Objektivierung der Notengebung

Die zweite Begründung für die Einführung von Leistungstests setzt bei der Notengebung an. Die Kritik an der Notengebung, die an sich nichts Neues darstellt, hat durch internationale Schulleistungsvergleichsstudien wie PISA neuen Auftrieb gewonnen. Der Notengebung wird vorgeworfen, den Standards eines psychologischen Messverfahrens nicht zu genügen, d.h. weder reliabel (verlässlich) noch valide (gültig) zu sein. Und objektiv seien Noten schon gar nicht. Allein schon die Messkriterien, wie sie durch Lehrpläne vorgegeben würden, seien zu unbestimmt, würden zu viel der Interpretation anheimstellen und liessen sich nicht operationalisieren. Von Lehrer zu Lehrer, von Klasse zu Klasse und von Schule zu Schule falle die Benotung ein und derselben Schülerleistung verschieden aus.[1] Die EDK (2006) spricht von «subjektiven und heterogenen Bewertungen von Schülerleistungen» (S. 24). Was Schülerinnen und Schüler *tatsächlich* können, bleibe offen (vgl. EDK 2004, S. 5).

Dabei machen es sich die Kritiker oft leicht, wie eine Expertise zeigen kann, die Jürgen Oelkers und Kurt Reusser (2008) für das (deutsche) Bundesministerium für Bildung und Forschung verfasst haben. Darin wird behauptet, Schulnoten würden «aus der Sicht der Lehrkraft erteilt» (S. 22) und einen «Klassendurchschnitt wieder(geben)» (ebd.), aber nicht «eine tatsächlich erreichte, durch materielle Fachstandards umschriebene Kompetenz (erfassen)» (ebd.). Die Noten seien «ohne Bezug auf ein fachinhaltlich-qualitatives, den Referenzrahmen eines zufällig zusammengesetzten Klassenzimmers übergreifendes Lernzielkriterium» (ebd.). «Das heisst: Noten beziehen sich auf den Durchschnitt einer Klasse, aber geben keinen Aufschluss über die in einem Fach oder einem Lerngebiet erreichten tatsächlichen Kompetenzen» (ebd.). Was die Schülerinnen und Schüler am Ende

[1] Dabei wird oft mit Studien argumentiert, die zeigen sollen, dass das Lehrerurteil den zentralen Ansprüchen an ein psychometrisches Messverfahren nicht genügt (vgl. Ingenkamp 1995). Bei den von Ingenkamp (1995) zusammengestellten und kommentierten Studien geht es jedoch zumeist um rein punktuelle, quasi-experimentelle Untersuchungen von Beurteilungsprozessen, die in keiner Weise berücksichtigen, dass die Induktionsbasis, die Lehrkräften für die Beurteilung von Schülerleistungen zur Verfügung steht, ausgesprochen breit ist und sowohl vergleichsweise große Zeiträume abdeckt als auch auf unterschiedlichen Methoden (Beobachtung, mündliche Daten, schriftliche Leistungen, Objektivationen verschiedener Art etc.) beruht. Die Qualität des Lehrerurteils ist bedeutend besser als sie in diesen Studien, deren ökologische Validität höchst fragwürdig ist, zum Ausdruck kommt.

eines Prozesses tatsächlich wissen und können, müsse «anders bestimmt werden» (ebd.). «Anders bestimmt werden» – das heisst: mittels externer Tests, wie Oelkers an anderer Stelle deutlich macht, wenn er schreibt, Tests seien «ein Mittel zur Objektivierung der Leistungsbeurteilung» (Oelkers 2007, S. 35).

Diese Art von Kritik der Notengebung spricht den Lehrkräften jegliche diagnostische Kompetenz ab. Es scheint jedoch dieser massiven Kritik zu bedürfen, damit als Alternative Leistungstests propagiert werden können. Der vermeintlich geringe Aussagewert von Noten soll mit Messinstrumenten kompensiert werden, die den Qualitätskriterien eines echten Messsystems genügen, und das sind standardisierte psychometrische Tests. Deren Vorteil liegt in der systematischen Entwicklung, die nicht nur garantiert, dass die Handhabung der Tests standardisiert erfolgt, sondern auch die Verlässlichkeit der Messung (Reliabilität) und der Aussagewert der Ergebnisse (Validität) garantiert werden können. Der Vorzug von psychometrischen Tests wird im Vergleich mit der Notenskala darin gesehen, dass sie die schulische (Leistungs-)Beurteilung von lokalen und subjektiven Normen ablösen und einer dekontextualisierten *Messung* unterwerfen lassen.

Mit der Notenskala steckt die Schule gewissermassen noch im 19. Jahrhundert, als nicht nur die Gewichte und Längenmasse, sondern auch die Uhrzeit noch von Ort zu Ort variierten (vgl. Messerli 1997). Inzwischen haben wir das 21. Jahrhundert erreicht, und es scheint nur angemessen zu sein, wenn sich endlich auch die Schulen einem metrischen Masssystem unterwerfen. Tests stehen für einen Modernisierungsschub, der die Schulen von lokalen und anthropomorphen Standards befreien und nationalen[2], ja internationalen und globalen Normen und Kriterien unterwerfen will.

1.3 Kontrolle und Steuerung

Ich komme zum dritten Argument, das zugunsten von Tests vorgebracht wird. Hier geht es nicht um Überwachung von Schule und Unterricht, auch nicht um Kritik an

[2] Die Klieme-Expertise bringt den nationalen Bezug bereits im Titel zum Ausdruck: «Zur Entwicklung nationaler Bildungsstandards» (Klieme et al. 2003). *HarmoS* beansprucht, die «Eckwerte» des Schulsystems schweizweit zu vereinheitlichen (EDK 2007).

der Messqualität von Noten, sondern um die Steuerung des Bildungssystems. Dieses soll wirksamer gesteuert werden als bisher.

Die traditionelle Form der politischen Steuerung ist die Steuerung über den Input. Jedes Gesetz, jede Verordnung und jedes Reglement hat steuernden Charakter, insofern es *vorschreibt,* was in einem bestimmten Bereich zu tun oder zu unterlassen ist. So will man das Verhalten der Banken steuern, indem man ihnen strengere Vorschriften macht, oder man will die Energiewende herbeiführen, indem man eine Ökosteuer einführt. Im gleichen Sinn kann man versuchen, die Schulen zu steuern, indem man festlegt, was eine Schule überhaupt auszeichnet, wie die Lehrer auszubilden sind und wie die Kinder und Jugendlichen das Schulsystem zu durchlaufen haben. Gesteuert wird über Vorschriften und Vorgaben, welche den Akteuren vor Ort Anweisungen geben, wie sie zu handeln haben.

Diese Art von Steuerung möchte man auswechseln. Das kann man unabhängig davon tun, ob man für oder gegen ein Bildungsmonitoring oder für oder gegen die Qualitätsverbesserung der Notengebung ist. Denn es geht um eine Managementphilosophie, d.h. um die Frage, wie ein privates oder staatliches Unternehmen zu führen ist, damit es seine Aufgaben und Dienstleistungen optimal erbringt. Das Stichwort lautet: New Public Management (NPM), zu deutsch: Neue Verwaltungsführung. Der Ansatzpunkt der Steuerung wechselt vom Input zum Output. Outputorientierte Steuerung meint, dass die Schulen nicht über Vorschriften und Vorgaben, sondern über die Beobachtung ihrer Erträge gesteuert werden. Vorgegeben wird nicht mehr, was die Schule, der Unterricht oder die Lehrkraft *tun* sollen, sondern, was sie *bewirken* sollen.

Offensichtlich sind auch dafür Tests notwendig. Denn ohne Tests, die überprüfen, was eine Schule tatsächlich leistet, ist es nicht möglich, ein Bildungssystem auf Outputsteuerung umzupolen. Wie Hermann Josef Abs (2009) zu Recht feststellt, wäre durch die bloße Proklamation von Leistungsstandards (was nichts anderes als ein weiterer Input wäre) «kein anderer Steuerungseffekt zu erwarten als von einer Lehrplanrevision. Erst die Verbindung von Leistungsstandards mit einer zentral organisierten Überprüfung der Leistungen in standardorientierten Tests ermöglicht eine andere Form des Regimes» (S. 827).

2. Test ist nicht gleich Test

Damit haben wir drei Begründungen für die Einführung von Tests an unseren Schulen gefunden: 1. Überwachung (Monitoring) des Bildungssystems, 2. Ausmerzung der Schwächen der Notengebung und 3. bessere Steuerung des Schulsystems. Interessanterweise lassen sich die drei Gründe unabhängig voneinander vertreten. Zwar werden Monitoring und Steuerung oft zusammen genannt, aber logisch gesehen besteht kein Zusammenhang – man kann ohne weiteres für ein besseres Monitoring sein, ohne zugleich auch für Outputsteuerung einzutreten. Und für die Verbesserung der Notengebung kann man sich stark machen, ohne die Verwendung von Tests für Monitoring- oder Steuerungszwecke für notwendig zu halten.

Hier liegt aber auch ein wesentlicher Grund, weshalb die Einführung von Tests zurzeit einen starken Rückhalt geniesst. Denn die verschiedenen Argumentationslinien überlagern sich und stützen sich gegenseitig. Wer für die Einführung einer Testkultur ist, dem kann es nur recht sein, wenn möglichst viele Gründe dafür sprechen. Denn so lässt sich leichter eine politische Mehrheit finden, als wenn es nur gerade *ein* Argument für Tests gäbe.

Der politische Konsens hat jedoch eine Kehrseite, die darin liegt, dass die wenigsten Verfechter einer schulischen Testkultur wissen, was ein Test überhaupt ist. Selbst Erziehungswissenschaftler, die sich für die Einführung einer schulischen Testkultur stark machen, sind oft wenig informiert über die Möglichkeiten, die psychometrische Tests bieten. Wie Daniel Koretz (2009, S. 327), einer der renommiertesten Psychometriker der USA, meint, sind unrealistische Erwartungen über Tests weit verbreitet. Dazu gehört, dass viele glauben, ein und derselbe Test lasse sich einsetzen, um allen drei zuvor genannten Zielen zu genügen. Doch dies ist eine schlichte Illusion. Es wird nicht möglich sein, mit ein und demselben Test sowohl die Notengebung zu verbessern wie auch das Bildungssystem zu überwachen und es auch gleich noch besser zu steuern. Test ist nicht gleich Test. Mit ein und demselben Test oder mit ein und derselben *Art* von Test kann nicht alles Beliebige erreicht werden. Ein Test ist keine Fliegenklappe, mit der man einmal zuschlagen und dreimal Erfolg haben kann.

Die Probleme, die mittels Tests gelöst werden sollen, sind in den drei genannten Fällen unterschiedlich. Das Monitoring ist auf der *Systemebene* angesiedelt. Es geht darum, dass wir wissen wollen, was unser Bildungssystem leistet. Das ist eine globale, ja abstrakte Problemstellung, bei deren Bearbeitung wir nicht ins Detail zu gehen brauchen. Die Messungen, die Tests im Rahmen eines Bildungsmonitoring oder einer externen Schulevaluationen erbringen, können relativ grob sein. Wir brauchen keine exakten Auskünfte, wenn wir erfahren wollen, wie unser Bildungssystem *im Grossen und Ganzen* funktioniert. Das heisst auch, dass Tests, die zum Zweck der Systembeobachtung eingesetzt werden, nicht flächendeckend zum Einsatz kommen müssen. Es genügt, wenn wir Stichproben bilden, die für einen Schultyp oder eine Schulstufe repräsentativ sind. Auch die Testauswertung kann auf der Aggregatebene verbleiben, d. h. von Aussagen über einzelne Klassen oder einzelne Lehrkräfte lässt sich genauso absehen, wie auf Urteile über einzelne Schülerinnen und Schüler verzichtet werden kann. Bei PISA und anderen internationalen Schulleistungsvergleichsstudien wird genau so vorgegangen, was es erlaubt, komplexe Stichprobenpläne zu nutzen, wie das sogenannte *matrix sampling*, bei dem den untersuchten Probanden unterschiedliche Subtests vorgelegt werden (vgl. Koretz 2009, S. 64f.).

Anders als die Leistungsprüfung des Bildungssystems, bei dem die Qualitätsansprüche an die Testqualität vergleichsweise bescheiden sein können, stellt die Objektivierung der Notengebung hohe Ansprüche an die Testqualität. Denn hier ist die Referenz nicht die abstrakte Systemebene, sondern die konkrete *Individualebene* der einzelnen Schülerinnen und Schüler. Der Messfehler, mit dem jeder psychologische Test behaftet ist und der bei jeder Messung ins Gewicht fällt, wird nicht mehr durch die Aggregierung der Daten im Rahmen einer Stichprobe von Schülerinnen und Schülern statistisch ausgeglichen, sondern schlägt voll auf das Individuum durch. Tests zu individualdiagnostischen Zwecken einzusetzen, ist daher nur legitim, wenn sie eine hohe Messgenauigkeit aufweisen. Das aber kann nur gewährleistet werden, wenn ein Test sorgfältig und verantwortungsvoll konstruiert wird, was nicht nur viel Geld kostet, sondern auch viel Zeit und viel Fachkompetenz in Anspruch nimmt. Wenn Tests mit der Etablierung einer schulischen «Förderkultur» in Verbindung gebracht werden, wie bei Oelkers (2007, S. 34),

dann stellen sich weit höhere Qualitätsansprüche, als wenn sie für das Systemmonitoring gedacht sind. Dasselbe gilt, wenn sie zu Selektionszwecken gebraucht werden. Es wäre völlig unangemessen, *denselben* Test *sowohl* für Monitoringzwecke *als auch* für Individualdiagnosen zu nutzen.

Irgendwo dazwischen liegt die Verwendung von Tests zu Steuerungszwecken. Allerdings ist genau dies nicht besonders klar. Einerseits könnte man sich mit Systemdaten begnügen, d. h. mit Daten von Tests, die beim Systemmonitoring anfallen. Andererseits stellt sich die Frage, wie wirksam eine Steuerung ist, die auf der Systemebene verbleibt. So abstrakt das System ist, so diffus ist ein Steuerungsimpuls, der nicht über diese Ebene hinauswirkt. Naheliegender ist daher, dass die Steuerung tiefer ansetzt, zum Beispiel auf der Ebene der *Einzelschule*. Dann aber muss auch in diesem Fall gewährleistet sein, dass die Qualität der Tests ausreicht, um verlässliche Aussagen auf der Schulebene oder allenfalls auf der Unterrichtsebene zu ermöglichen. Um auf die einzelne Schule steuernd Einfluss zu nehmen, ist zwar nicht dieselbe hohe Datenqualität nötig, wie sie Individualdiagnosen voraussetzen, aber doch eine höhere, als wenn es lediglich um das Systemmonitoring geht.

3. *High-Stakes Testing* in den USA

Die Unterscheidung verschiedener Zwecke, wofür Tests eingesetzt werden, zeigt, dass die Kritik der Notengebung keine guten Karten hat, wenn sie glaubt, die Einführung einer Testkultur, wie sie von der Standardbewegung propagiert wird, lasse die Qualität der Schülerbeurteilung verbessern. Es ist zwar ohne weiteres möglich, Tests zu konstruieren, die sich für Individualdiagnosen eignen, und dies auch im Bereich von Schule und Unterricht. Aber der Aufwand, der dabei zu leisten ist, ist beträchtlich, auch und gerade in finanzieller Hinsicht. Und er übersteigt vermutlich die realen Möglichkeiten, die ein Bildungssystem hat, um sich zu optimieren.

Ich möchte dies am Beispiel der USA illustrieren, da wir dort eine weit entwickelte Testkultur vorfinden, die in jüngster Zeit aufgrund politischer Einflussnahmen jedoch zu gravierenden Problemen geführt hat. Um zu zeigen, wo das Hauptpro-

blem liegt, ist ein kurzer Blick in die Geschichte der amerikanischen Bildungspolitik notwendig.

Man kann die Bildungspolitik in den USA in Bezug auf die letzten rd. 30 Jahre etwas grobschlächtig in drei Phasen einteilen:

1980 – 1990: Testbewegung *(testing movement)*
1990 – 2000: Standardbewegung *(standard movement)*
2000 – heute: Rechenschaftsbewegung *(accountability movement)*

Man kann es aber auch so sehen, dass diese drei Phasen Subphasen einer *einzigen* Bewegung sind, die sich insgesamt Standardbewegung nennt. Dafür spricht, dass die standardbasierte Schulreform in den USA zumeist mit dem Bericht einer Kommission der Reagan-Administration in Verbindung gebracht wird: *A Nation at Risk* (1983). Standards spielen in diesem Bericht keine nennenswerte Rolle, und wenn, sind nicht Leistungsstandards (performance standards), sondern curriculare Standards (content standards) gemeint. Im Vordergrund stehen Empfehlungen herkömmlicher Art, und das bedeutet im Kontext der USA nicht zuletzt ein Ausbau des Einsatzes von Tests zur besseren Überwachung des Bildungssystems.

George Madaus, Michael Russell und Jennifer Higgins (2009, S. 19) sprechen von einer «tektonischen Verschiebung» (tectonic shift), die dieser Bericht der Reagan-Administration in Bezug auf den Einsatz von Tests im amerikanischen Schulwesen ausgelöst hat. Die Autoren schreiben: «As seen in previous decades, policy makers in the 1980s relied on test scores to argue that there was a problem in our educational system. Policy makers then called for increased testing in order to determine whether reforms were effective. As the 1980s drew to a close, testing was widely accepted as an essential tool for improving education» (ebd., S. 20).

Die 1980er-Jahre waren demnach eine Zeit der Expansion von Tests im amerikanischen Schulwesen (deshalb der Ausdruck *testing movement*). Eine Qualitätsverbesserung der Schulen oder eine Steigerung der Schülerleistungen hat jedoch nicht stattgefunden. Das löste Anfang der 1990er-Jahre die *Standardbewegung* im engeren Sinn des Wortes aus. Die 1990er-Jahre waren die Zeit der Präsidentschaft von Bill Clinton. Unter Clinton wurde vor allem versucht, die Lehrpläne zu

vereinheitlichen. Genau dafür verwendete man den Begriff der Standards, d. h. nicht in dem Sinne, wie der Begriff bei uns in erster Linie verwendet wird, nämlich als Standards für Schülerleistungen *(performance standards),* sondern im Sinne von curricularen Standards *(curricular bzw. content standards).*

Die Standardbewegung der 1990er-Jahre ist im Wesentlichen gescheitert. Und zwar weil es nicht gelang, sich auf gemeinsame inhaltliche Standards zu einigen, wofür das Fach Geschichte den exemplarischen Fall darstellt. Diane Ravitch (2010) schreibt in ihrem Buch «The Death and Life of the Great American School System»: «The standards movement died in 1995, when the controversy over the national history standards came to a high boil. And the state standards created as a substitute for national standards steered clear of curriculum content. So, with a few honorable exceptions, the states wrote and published vague documents and called them standards. Teachers continued to rely on their textbooks to determine what to teach and test. The tests and textbooks, written for students across the nation, provided a low-level sort of national standard. Business leaders continued to grouse that they had to spend large amounts of money to train new workers; the media continued to highlight the mediocre performance of American students on international tests; and colleges continued to report that about a third of their freshmen needed remediation in the basic skills of reading, writing, and mathematics» (S. 20). Das heisst nichts anderes, als dass auch dieser (zweite) Reformansatz gescheitert ist. Und damit kommt, was bei *uns* im Allgemeinen unter standardbasierter Schulreform verstanden wird, nämlich ein Wechsel von der Steuerung des Schulsystems über den Input zu einer Steuerung über den Output. Dafür steht in den USA die Bezeichnung *accountability-movement*.

Anfang 2001 wird Bush jr. Präsident der USA. Die Bildungspolitik bezeichnete er als eine seiner Prioritäten, was sich im sogenannten *No-Child-Left-Behind*-Gesetz konkretisierte, das im Herbst 2001 mit grossen Mehrheiten im Kongress und im Repräsentantenhaus verabschiedet und von George W. Bush am 8. Januar 2002 in Kraft gesetzt wurde. Das Gesetz verlangt strikte Kontrollen der Schulen, wobei die Kontrollen mit Sanktionen verbunden sind. Dafür steht nicht nur der Begriff der

accountability, der mit Rechenschaftsgabe übersetzt werden kann[3], sondern auch derjenige des *high-stakes testing*.[4]

High-Stakes Testing ist die Verwendung von Tests – irgendwelcher Tests – zum Zweck der Entscheidungsfindung in Bezug auf Einzelfälle, wobei die Einzelfälle Schulen oder Schulleitungen, Lehrpersonen oder Schülerinnen bzw. Schüler sein können (vgl. Gamoran 2007, S. 82ff.). Die Entscheidungen haben für die betroffenen Personen unter Umständen gravierende Folgen, wie zum Beispiel im Falle eines Schülers die Repetition einer Klasse, die Überweisung an eine Sonderschule oder die Verweigerung eines Abgangszeugnisses, im Falle eines Lehrers eine Gehaltskürzung oder die Kündigung und im Falle einer Schule deren Verpflichtung, Massnahmen zur Verbesserung der Schülerleistungen zu treffen, die Setzung der Schule «auf Bewährung» (*on probation*), die Entlassung der Schulleitung, die Privatisierung der Schule (z.B. als *Charter School*) oder die Berechtigung der Eltern, ihr Kind von der Schule abzuziehen. Die Verwendung von Tests für *High-Stakes Decisions* ist in den Augen von Koretz (2009) «beyond a doubt, the single most important change in testing in the past half century» (S. 57) in den USA.

Damit sehen wir, dass das wirklich Neue an der standardbasierten Schulreform der Zugriff auf die tieferen Ebenen des Schulsystems ist. Von der politischen bzw. administrativen Steuerung des Bildungssystems wird eine Wirksamkeit erwartet, die über die Systemebene hinaus reicht und sich letztlich bis auf die Individualebene der Schülerinnen und Schüler, zumindest aber der Lehrerinnen und Lehrer erstreckt, die offensichtlich für die Schülerleistungen verantwortlich gemacht

[3] Eine Definition gibt Winch (1996): «An organization is accountable if it is possible to determine whether or not it fulfills the purposes for which it was set up. So accountability can be exercised by checking to see whether it is fulfilling its purposes and *how well* it is fulfilling them» (S. 33). Ähnlich heisst es bei Ravitch (2010): «By accountability, elected officials said that they wanted the schools to measure whether students were learning, and they wanted rewards or punishments for those responsible» (S. 95). Mit *Accountability* ist also gemeint, dass überprüft wird, ob eine Schule die ihr gesetzten Ziele erreicht, und dass sie zur Verantwortung gezogen wird, falls dies nicht der Fall sein sollte. Insofern steht die Rechenschaftspflicht im Interesse der besseren (Output-)Steuerung des Bildungssystems.

[4] Statt von *high-stakes testing* ist auch der Begriff des *accountability testing* geläufig (vgl. Boardman & Woodruff 2004), was unmittelbar zum Ausdruck bringt, dass mit den Testergebnissen eine Rechenschaftspflicht bzw. Haftbarkeit einhergeht.

werden sollen. Dabei dienen die Tests gleichsam als Scharnier, um die verschiedenen Ebenen des Schulsystems miteinander zu verbinden. Den Tests wird nicht mehr nur (wie bisher) eine Monitoringfunktion, sondern eine Steuerungsfunktion zugewiesen, und zwar nicht auf der Systemebene, sondern – wie gesagt – letztlich auf der Individualebene, zumindest aber auf der Schul- und Unterrichtsebene.

4. Die Situation in der Schweiz

Wenn dies die Situation in den USA ist, wie weit davon entfernt sind wir dann hierzulande? Die Frage ist nicht leicht zu beantworten. Einerseits vernehmen wir Voten, wonach Tests, die zu Monitoringzwecken entwickelt werden, nicht für Förder- oder Selektionszwecke eingesetzt werden sollen. In der Klieme-Expertise beispielsweise wird deutlich darauf hingewiesen, dass sich Tests, mittels derer Bildungsstandards überprüft werden, *nicht* für Individualdiagnosen eignen (vgl. Klieme et al. 2003, S. 10). Von der Verwendung von Tests für individualdiagnostische Urteile wird sogar explizit abgeraten (ebd., S. 49, 107ff.). Ähnliches hört man von Seiten der EDK. Mit Bildungsstandards, so wird uns versichert, sollen «in keiner Weise Leistungen von Schulen … verglichen oder ein entsprechendes Ranking erstellt werden» (Maradan, 2007, S. 99). Angestrebt würde allein, «die Leistungen des Schulsystems zu überprüfen und allenfalls die Wirkung einer Schulreform präzise zu dokumentieren» (Mangold, Rhyn & Maradan, 2005, S. 179). Das heisst, man will sich auf die *Systemebene* und das *Systemmonitoring* beschränken und allenfalls noch die Ebene der Einzelschule einbeziehen, aber nur zu Evaluationszwecken, nicht für Einzeldiagnosen.

So klar diese Sätze sind, so wenig eindeutig ist die Situation. Sobald die Politik mitredet, hört man auch andere Töne. Die (deutsche) KMK schreibt in ihrem Beschluss vom 2. Juni 2006: «Neben ihrer Funktion der Beschreibung von Leistungsanforderungen und der Leistungsmessung dienen die Bildungsstandards primär der Weiterentwicklung des Unterrichts und vor allem [!] der verbesserten individuellen Förderung aller Schülerinnen und Schüler» (KMK 2006, S. 13). Offenbar sollen die Tests doch *auch* für Entscheidungen auf der Individualebene genutzt werden.

Genauso ambivalent ist die EDK. In ihrem Vernehmlassungsbericht zum *HarmoS*-Projekt heisst es, der den Bildungsstandards zugrunde liegende Referenzrahmen werde «auch für die Entwicklung bzw. Anpassung von Instrumenten zur *individuellen Standortbestimmung der Schülerinnen und Schüler* verfügbar sein» (EDK, 2006, S. 23 – Hervorhebung W.H.). HarmoS werde sogar erlauben, «durch standardisierte Tests die Schülerbeurteilung zu *verbessern*» (ebd., p. 26 – Hervorhebung W.H.). Das heisst nichts anderes, als dass die Tests, die für das Systemmonitoring eingesetzt werden, auch für Förderzwecke zum Einsatz kommen sollen.[5]

Diese Position hat die EDK erst kürzlich wieder bekräftigt. In der neuesten Nummer ihres Publikationsorgans éducation[ch] heisst es, dass nun Testaufgaben entwickelt würden, die auf gesamtschweizerischer Ebene zur Überprüfung der Bildungsstandards zum Einsatz kämen. «Für unsere Schulkultur» würde aber leitend bleiben: «Keine Testitis und keine Rankings» (S. 1). Die EDK verfolge ein Evaluationskonzept, das bewusst zurückhaltend sei. Auf gesamtschweizerischer Ebene würden die Tests für das Systemmonitoring eingesetzt. «Dieses umfasst repräsentative Stichproben. Rankings oder Beurteilungen von Lehrpersonen sind damit nicht möglich und auch nicht gewollt» (ebd.). Dann aber heisst es wiederum: «Eine zweite Anwendung finden die Tests im Rahmen der förderorientierten individuellen Standortbestimmungen, die in der Verantwortung der Sprachregionen entwickelt werden» (ebd.). Erneut wird die Systemebene mit der Individualebene kurzgeschlossen, was die Frage aufwirft, wie dies bewerkstelligt werden soll.

Denn entweder werden die Tests für das Bildungsmonitoring entwickelt, dann wird ihre Qualität nicht ausreichen, um individuelle Diagnosen zu stellen. Oder sie lassen Individualdiagnosen zu, dann sind sie von einer Qualität, die es unwahrscheinlich erscheinen lässt, dass man sie nicht auch für ein Lehrerranking oder ein Schulranking einsetzen wird. Die völlig unklare Position der EDK lässt den Verdacht aufkommen, dass man die Lehrer damit ködern will, indem man ihnen Instrumente für die Verbesserung der Schülerbeurteilung in Aussicht stellt, während

[5] Noch weiter geht die Erwartung, Bildungsstandards könnten einen Beitrag zur Einlösung der Chancengleichheit an unseren Schulen leisten. Behauptet wird gar, «erst [!] mit dem HarmoS-Projekt» würden «die Grundlagen … für eine Verbesserung der Bildungschancen für Kinder und Jugendliche mit benachteiligter sozialer Herkunft geschaffen» (EDK, 2008, S. 43).

es letztlich – wie die Situation in den USA zeigt – um eine Verschärfung der Kontrolle geht.

Ich komme zum Schluss keiner Ausführungen. Was sich standardbasierte Schulreform nennt, ist die Überzeugung, das Schulsystem lasse sich bis auf die Ebene des Schülerlernens von oben herab durchsteuern. Wobei Tests das Instrument bilden, um diesen Durchgriff zu ermöglichen. Weshalb sollte es bei uns anders kommen, als es inzwischen in den USA gekommen ist? Sobald es jedoch dazu kommt, verlieren die Tests ihre Messqualität. Die den Lehrkräften zugewiesene Inkompetenz in Sachen Notengebung kehrt als unerwünschte Nebenwirkung einer extensiven Testanwendung zurück. Auch dies kann man am Beispiel der USA sehr genau beobachten.

Im Vorwort zu einem Bericht über die grossen Unterschiede zwischen 26 Staaten der USA bei der Festlegung dessen, welche Schülerleistungen im Lesen und in Mathematik – entsprechend den Forderungen des *No-Child-Left-Behind*-Gesetzes – als *proficient* gelten[6], schreiben Finn und Petrilli: «... the testing enterprise is unbelievably slip shot. It's not just that results vary, but that they vary almost randomly, erratically, from place to place and grade to grade and year to year in ways that have little or nothing to do with true differences in pupil achievement. ... The testing infrastructure on which so many school reform reports rest, and in which so much confidence has been invested, is unreliable – at best» (ebd., S. 3). Das ist eine überraschende Aussage. Denn sind dies nicht genau die Vorwürfe, die hierzulande an die Adresse der Lehrer und die *Notengebung* formuliert werden? Hier, in dem Zitat von Finn und Petrilli, werden sie gegenüber den *Tests* erhoben! Dass es mit der flächendeckenden Einführung von Tests und dem «Einstieg in eine Testkultur» besser wird an unseren Schulen, scheint demnach eine illusorische Erwartung zu sein.

[6] Der Bericht kommt zum Ergebnis, dass die Bandbreite des *cut score*, der festlegt, was als *proficient* gilt, auf einer Perzentilskala vom 6. bis zum 77. Perzil reicht. Verschiedene Staaten haben ihr *proficient*-Niveau im Verlaufe der vergangenen Jahre nach unten korrigiert. «The whole rational for standards-based reform was that it would make expectations for student learning more rigorous and uniform. Judging by the findings of this study, we are as far from that objective as ever» (Cronin, Dahlin, Adkins & Kingsbury 2007, S. 7).

Damit klärt sich die eingangs noch offen gelassene Deutung meines Referatstitels: Eine Testkultur an unseren Schulen einzuführen, dies hat uns *gerade noch* gefehlt!

Literaturverzeichnis

Abs, H. J. (2009). Standards schulischer Bildung. In S. Andresen, R. Casale, Th. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 819-833). Weinheim: Beltz.

Boardman, A. G. & Woodruff, A. L. (2004). Teacher Change and «High-Stakes» Assessment: What Happens to Professional Development? *Teaching and Teacher Eduaction, 20*, 545-557.

Cronin, J., Dahlin, M., Adkins, D. & Kingsbury, G. G. (2007). *The Proficiency Illusion*. Washington, DC: Thomas M. Fordham Institute & Northwest Evaluation Association.

EDK [Schweizerische Konferenz der kantonalen Erziehungsdirektoren] (2004). *HarmoS. Zielsetzungen und Konzeption. Weissbuch.* Bern: EDK.

EDK [Schweizerische Konferenz der kantonalen Erziehungsdirektoren] (2006). *Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule. HarmoS-Konkordat. Bericht zur Vernehmlassung*. Bern: EDK.

EDK [Schweizerische Konferenz der kantonalen Erziehungsdirektoren] (2007): *Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule vom 14. Juni 2007*. Bern: EDK.

EDK [Schweizerische Konferenz der kantonalen Erziehungsdirektoren] (Hrsg.) (2008). *Lehrberuf. Analyse der Veränderungen und Folgerungen für die Zukunft*. Bern: EDK.

Finn, Jr., C. F. & Petrilli, M. J. (2007). Foreword. In J. Cronin, M. Dahlin, D. Adkins & G. G. Kingsbury, *The Proficiency Illusion* (S. 2-5). Washington, DC: Thomas M. Fordham Institute & Northwest Evaluation Association.

Gamoran, A. (2007). School Accountability, American Style: Dileammas of High-Stakes Testing. *Schweizerische Zeitschrift für Bildungswissenschaften, 29*, 79-94.

Herzog, W. (2010). Besserer Unterricht dank Bildungsstandards und Kompetenzmodellen? In A. Gehrmann, U. Hericks & M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 37-46). Bad Heilbrunn: Klinkhardt.

Herzog, W. (2011). Professionalität im Beruf der Lehrerinnen und Lehrer. In H. Berner & R. Isler (Hrsg.), *Lehrer-Identität, Lehrer-Rolle, Lehrer-Handeln* (S. 49-77). Baltmannsweiler: Schneider Verlag Hohengehren.

Ingenkamp, K. (Hrsg.) (1995). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte* (9. Aufl.). Weinheim: Beltz.

Klieme, E. et al. (2003): *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bonn: Bundesministerium für Bildung und Forschung.

KMK [Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland] (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Berlin: KMK.

Koretz, D. (2009). *Measuring Up. What Educational Testing Really Tells Us*. Cambridge, Mass.: University of Harvard Press.

Madaus, G., Russell, M. & Higgins, J. (2009). *The Paradoxes of High Stakes Testing. How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*. Charlotte, NC: Information Age Publishing.

Mangold, M., Rhyn, H. & Maradan, O. (2005). Leistungsstandards (HarmoS) und Bildungsmonitoring: zwei Hauptprioritäten der EDK und die Funktion der externen Evaluation. In G. Brägger, B. Bucher & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 175-185). Bern: h.e.p.

Maradan, O. (2007). Bildungsstandards in der Schweiz. In P. Labudde (Hrsg.), *Bildungsstandards am Gymnasium. Korsett oder Katalysator?* (S. 97-104). Bern: h.e.p.

Oelkers, J. (2007). Bildungsstandards im Gymnasium – Ein neues Problem? In P. Labudde (Hrsg.), *Bildungsstandards am Gymnasium. Korsett oder Katalysator?* (S. 27-36). Bern: h.e.p.

Oelkers, J. & Reusser, K., unter Mitarbeit von E. Berner, U. Halbheer und S. Stolz (2008). *Qualität entwickeln – Standards sichern – mit Differenz umgehen*. Bonn: Bundesministerium für Bildung und Forschung.

Ravitch, D. (2010). *The Death and Life of the Great American School System. How Testing and Choice Are Undermining Education*. New York.

Winch, Chr. (1996). *Quality and Education*. (Journal of Philosophy of Education: Special Issue.) Oxford: Blackwell.