Bildungsevaluation in einem Schweizer Kanton– eine kritische Nachbetrachtung

Kurzgutachten

Konstantin Beck
CSS-Institut für empirische Gesundheitsökonomie[1]
www.css-institut.ch

Luzern, im November 2009

1. Die Qualität schulischer Leistungen

Das Messen und Vergleichen von schulischen Leistungen ist in jüngster Zeit, vor allem seit der Popularisierung der Pisa-Resultate auch in der Schweiz im Vormarsch. Dahinter steht die Idee, die Qualität der Bildung fassbar zu machen. Längerfristiges ist es sicher auch das Ziel, den Eltern eine Entscheidungsgrundlage bei der Wahl zwischen konkurrierenden Schulen zu liefern.

Die Messung der Schulqualität dürfte daher mittelfristig erhebliche Auswirkungen auf die Bildungslandschaft der Schweiz haben. Zu den ersten Pilotversuchen gehört auch die hier zu diskutierende Vergleichsprüfung in Mathematik 2009, welche in 12 Mittelschulklassen mit 239 Schülerinnen und Schülern eines Schweizer Kantons durchgeführt worden ist (Keller & Moser, 2009).

Alle Schülerinnen lösten dieselbe aus 11 Teilfragen bestehende Mathematik-Prüfung. Die Prüfungen wurden anschliessend von einer Lehrperson korrigiert, welche die jeweilige Klasse nicht unterrichtet hatte. Die pro Klasse resultierende mittlere Punktzahl wurde zur Vergleichsgrösse. Hinter dieser Evaluation steckt die Idee, dass das durchschnittliche Abschneiden der Schüler einer Klasse die Qualität des jeweiligen Mathematikunterrichts widerspiegeln soll.

Diesem Kurzgutachten liegen die individuellen Resultate der Schüler und Schülerinnen aus zwei der 12 verglichenen Klassen zu Grunde. Es geht hier nicht darum, der Frage nach zu gehen, inwiefern das durchschnittliche Abschneiden der Schüler bei einer Mathematikprüfung mit der Qualität der unterrichtenden Lehrperson in direkte Verbindung gebracht werden kann. Immerhin dürften auch andere Einflussfaktoren eine Rolle spielen, die bei dieser Versuchsanlage nicht auskorrigiert werden. Beispielsweise

[1] Kontakt: CSS Institut für empirische Gesundheitsökonomie / Tribschenstrasse 21 / CH-6002 Luzern / ++41 058 277 12 73 / konstantin.beck@css.ch

kann das mittlere Prüfungsergebnis auch Folge einer Selbstselektion des jeweiligen Schülerkollektivs sein etc.

Es geht ebenso wenig darum, eine möglicherweise fehlende Feinabstimmung und Homogenisierung der Bewertungsmassstäbe der verschiedenen beteiligten Lehrpersonen bei der Korrektur unvollständig gelöster Aufgaben ins Feld zu führen. Auch soll nicht auf das Fehlen eines gemeinsamen Lehrplans im Bereich der Mittelschulmathematik im betrachteten Kanton eingegangen werden, obwohl das eine wichtige Voraussetzung für einen Leistungsvergleich unterschiedlicher Schulklassen darstellt.

Hier interessiert primär die Frage, ob die so gewonnen Resultate statistisch korrekt interpretiert werden. Oder anders ausgedrückt, ob die durchgeführte Evaluation aus statistisch formaler Sicht überhaupt in der Lage ist, Qualitätsunterschiede zwischen den Lehrpersonen auszumachen. Keller und Moser diskutieren diese Frage nicht ansatzweise und machen sich die Sache ein wenig einfach, wenn sie schreiben: „Diese Beschreibungen enthalten allerdings keine Beurteilung der Ergebnisse. Die Beurteilung der Ergebnisse können Sie als Lehrperson am zuverlässigsten vornehmen, weil Sie die Lernvoraussetzungen der Schülerinnen und Schüler am besten kennen."

2. Das Problem der Rangtransformation – der Dagobert Duck-Effekt

Sobald Messresultate vorliegen, gehen Medien und Publikum dazu über, eine Rangreihenfolge abzuleiten, d.h. eine Rangtransformation vorzunehmen. Rangtransformationen haben den Vorteil, dass sie die relative Position einer Leistung oder eines gemessenen Attributs stärker betonen, während sie die absoluten Unterschiede zwischen den Attributen eher verwischen.

Aus dem Sport, wo tagtäglich beispielsweise Laufzeiten in Ranglisten transformiert werden, sind uns Hinweise der Sportreporter folgender Art geläufig: Die ersten 10 Fahrer haben sich innerhalb einer einzigen Sekunde platziert, oder umgekehrt, die ersten zwei Fahrer hätten sich um mehrere Sekunden vom Feld distanziert. Die Information des einzelnen Rangs wird hier als ungenügend empfunden, weil sie die Knappheit respektive Deutlichkeit des Resultats nicht mehr widerspiegelt. Da wir im Sport grosse Übung mit Rangtransformationen haben, wissen wir intuitiv, dass eine Tabellenführung im Fussball mit einem Punkt Abstand zwar einen virtuellen ersten Platz bedeutet, jedoch schon durch ein Unentschieden des Zweitplazierten möglicherweise wett gemacht werden könnte.

Auch Reportagen, die bei einem Zweikampf vom ausgezeichneten zweiten Platz und von der blamablen zweitletzten Platzierung zu berichten wissen, werden die wenigsten hinters Licht führen. Wenn wir aber die vertraute Welt des Sports verlassen, und uns über Ranglisten informieren, die neu sind und deren Messung uns nicht im selben Ausmasse vertraut ist, können wir über folgenden Dagobert Duck-

Effekt stolpern. Während das Einkommen des reichsten Mannes Entenhausens dasjenige seines Neffen Donald um Fantasilliarden übersteigt, schrumpft der Abstand durch die Rangtransformation in bescheidene Grössen. Innerhalb der engeren Familie Duck belegt Dagobert bezüglich der Einkommen den ersten und Donald den zweiten Rang. Plötzlich erscheinen die beiden Duck-Protagonisten nahezu auf Augenhöhe.

Alle diese Beispiele machen eines deutlich: Ranglisten stellen transformierte Zahlenreihen dar, wobei die Rangtransformation wichtige, das Bild abrundende Tatsachen unterschlagen kann.

3. Unzulässige Ranglisten – der inverse Dagobert Duck-Effekt

Bisher diskutierten wir lediglich irreführende Effekte von Rangtransformation. In diesem Abschnitt geht es um schwerwiegendere Fehler der Rangtransformation. Es gibt Fälle, wo Rangtransformationen gar nicht zulässig sind. Das sind die Fälle, bei denen Rangunterschiede suggeriert werden, die in Wahrheit nicht vorliegen. Wird beim Dagobert Duck-Effekt ein enormer Unterschied miniaturisiert, so wird beim inversen Effekt ein nicht existenter Unterschied aufgebläht.

Im Sport mag zwar der Zufall im Spiel gewesen sein, wenn die eine Fahrerin eine andere um eine Hundertstel-Sekunde vom Podest verdrängt. Dennoch ist der Fall klar. Gemäss Reglement gebührt der Schnelleren von beiden die Auszeichnung. Ein Rennen ist keine Stichprobe, die den Anspruch auf Allgemeingültigkeit erhebt. Einzig die gemessene Zeit ist entscheidend.

Ganz anders ist das bei Indikatoren, die, wie im vorliegenden Fall, Auskunft über die Qualität eines Schulunterrichtes oder gar einer Schule liefern sollten. Diese Indikatoren sind nicht einfach und schon gar nicht präzise messbar. Zufälligkeiten können eine grosse Rolle spielen. Solche Effekte sind im Auge zu behalten und bei der Publikation von Resultaten entsprechend anzuführen.

Es gibt eine alte und reichhaltige statistische Tradition, wie mit zufallsbehaftetem Datenmaterial zu verfahren sei. Die Statistik kennt dafür den Begriff der Signifikanz. Mit statistischen Formeln ist es möglich, zwischen zufälligen und systematischen Messwert-Unterschieden zu unterscheiden. Liegen nun Zahlen aus Messungen vor, die sich nicht systematisch sondern nachweisbar nur zufällig voneinander unterscheiden, so dürfen diese Resultate *nicht* in Rangfolgen transformiert werden. Denn genau so wie die Messresultate sind dann auch die Rangierungen zufällig. Das kann so weit gehen, dass Eltern ihre Kinder aus der einen Schule nehmen und in eine andere Schule schicken, obwohl sich die beiden Schulen in Tat und Wahrheit nicht messbar unterscheiden. Oder das staatliche Mittel der einen Schule vorenthalten bleiben, obwohl sie eine vergleichbare Ausbildungsqualität aufweist, wie die subventionierten Schulen.

Wenn man weiter in Betracht zieht, dass das Bilden von Ranglisten gerade in jüngster Zeit – gefördert von einer auf Ranglisten fixierten medialen Öffentlichkeit – unglaublich populär ist, dann liefern die Qualitätsmessungen der Mathematikleistungen unterschiedlicher Mittelschulen im betrachteten Kanton ein äusserst problematisches Beispiel einer solchen unzulässigen Rangtransformation. Sind diese Werte einmal veröffentlicht, werden verschiedene Exponenten unweigerlich daraus eine Rangfolge ableiten, nicht wissend, dass diese Rangierung äusserst fragwürdig ist.

Auch die beiden Autoren selbst sprechen deutlich von der besten und der schwächsten Klasse. Warum eine solche eindeutige Terminologie fehl am Platz ist, soll im Folgenden gezeigt werden.

4. Datenbasis

In einem Schweizer Kanton wurden die Mathematikprüfungen von 239 Schülerinnen und Schülern aus 12 Klassen untersucht (das ergibt im Schnitt 19.9 Schüler pro Klasse). Für diese Nachanalyse lagen die detaillierten Ergebnisse von zwei Klassen (mit 23 respektive 17 Schülern) vor. Ferner waren die Durchschnittlichen Punktzahlen für vier Klassen exakt bekannt und für die übrigen 8 Klassen auf Grund einer Graphik auf etwa 0.1 Punkt rekonstruierbar. Tabelle 1 zeigt die exakten und auf Grund der Graphik geschätzten Klassenmittelwerte:

Tabelle 1: Mittlere Punktzahl pro Klasse (geschätzte Werte *kursiv*)

Klasse Nr.	Mittlere Punktzahl
1	12.7
2	12.5
3	12.4
4	*12.3*
5	*11.5*
6	*11.4*
7	*10.8*
8	*10.6*
9	*10.5*
10	*10.2*
11	*10.0*
12	7.9

Aus dem Text wissen wir, dass die erreichten Punktwerte, die zwischen 1 und 24 streuen, linkssteil verteilt sind, mit einem Modus bei etwa 10 Punkten.

5. Keine Schule in diesem Kanton ist die Beste

Bei zwei Klassen (A und B) waren die detaillierten Resultate verfügbar. Dabei fiel auf, dass in Klasse A (mit 23 Schülern) sowohl das globale Maximum (24 Punkte) als auch das globale Minimum (1 Punkt) auftrat. D.h., dass diese Klasse einen sehr hohen Durchschnitt aufwies, obwohl die schwächste Schülerin zum Klassenkollektiv gehörte. Es zeigt sich dadurch aber auch sofort eine Manipulationsgefahr: Hätte der betreffende Lehrer diese Schülerin krankschreiben lassen, dann läge seine Klasse mit 13.0 Punkten an der Spitze. Dasselbe Argument trifft für Klasse B zu, die ohne ihren schwächsten Schüler ebenfalls den Spitzenplatz mit 12.9 Punkten belegt hätte.

Umgekehrt hätte das Fernbleiben des besten Schülers Klasse A drei Ränge und Klasse B zwei Ränge gekostet. Das illustriert bereits eine gewisse Zufallsabhängigkeit des gemessenen Resultats, die entsprechenden Schüler hätten ja auch tatsächlich krank sein können.

Die Standardabweichung der grösseren Klasse betrug 5.8, die der kleineren 5.4. Die Standardabweichung des gemessenen Mittelwerts reduziert sich um die Wurzel der Schülerzahl, also für Klasse A: $5.8 / 20^{1/2} = 1.21$. Für Klasse B resultiert 1.31.

Es ist auf Grund der Verteilung der 239 Punktwerte nicht unplausibel anzunehmen, dass die Klassenmittelwerte nahezu normalverteilt sind. Trifft diese Annahme zu, so besagt eine Standardabweichung von 1.21 folgendes: Mit 68,3% Wahrscheinlichkeit streuen die Messresultate dieser Klasse plus/minus eine Standardabweichung um den Mittelwert. Konkret: Mit gut 2/3 Wahrscheinlichkeit ist mit dem zufälligen Auftreten von Messresultaten der Klasse A im Intervall $12.5 + 1.21 = \mathbf{13.71}$ und $12.5 – 1.21 = \mathbf{11.29}$ zu rechnen. Mit Blick auf Tabelle 1 zeigt sich, dass sich das Resultat der Klasse A nicht signifikant von den Rängen 1, 3, 4, 5 und 6 unterscheidet, weil alle diese Ränge im Intervall (13.71; 11.29) liegen.

68.3% ist ein unüblich geringes Signifikanzniveau. Erhöht man das Sicherheitsniveau auf 95,4% (respektive auf 2 Standardabweichungen) so lautet das Intervall, das zufällige Messresultate der Klasse A umschliesst: (14.95; 10.90). Auch in diesem Fall ist Klasse A nicht vom angeblich Bestklassierten bis zum an sechster Stelle Klassierten zu unterscheiden.

Noch deutlicher fallen die Resultate für Klasse B aus: Auf dem 68% Signifikanzniveau resultiert ein Intervall von (13.72; 11.10), d.h. die Resultate der Klasse B unterscheiden sich nicht signifikant von den Rängen 1, 2, 4, 5 und 6. Auf dem 95%-Niveau lassen sich kaum noch Unterschiede ausmachen. Nur noch der Schlussrang liegt ausserhalb der Zufallsstreuung der Messung für Klasse B (das Intervall lautet (15.03; 9.79)).

Fazit: Anstatt von 12 unterschiedlichen Rängen muss von elf identischen Mittelschulen und einer signifikant abfallenden Klasse gesprochen werden. Die Unterschiede auf den Rängen 1 bis 11 sind nicht wirklich von zufälligen Resultatschwankungen zu unterscheiden.

6. Empfehlungen für zukünftige Qualitätsmessungen

Die Messung der Bildungsqualität ist ein wichtiges aber auch schwieriges Unterfangen. Genau so viel Sorgfalt, wie der Messung der Qualitätsunterschiede gewidmet werden muss, muss auch auf die Präsentation der Ergebnisse verwandt werden.

Die vorliegende Analyse kommt zum Schluss, dass bei mindestens 11 von 12 Mittelschulen dieses Kantons die Qualitätsunterschiede rein zufälliger Natur sind und weiter reichende Schlussfolgerungen bezüglich dieser 11 Schulen irreführend weil zufallsbedingt sind. Wer nun aber bei der Präsentation der Schlussresultate jeden Hinweis auf die doch stark eingeschränkte Aussagekraft der Resultate vermissen lässt, und den wichtigen Aspekt statistischer Signifikanz ausser Acht lässt, trägt nicht nur zu falschen Schlussfolgerungen von erheblicher Tragweite bei, sondern erweist mittelfristig auch dem Wissenschaftszweig, den er vertritt, einen Bärendienst.

8. Literatur

Keller, Florian & Urs Moser (2009): "Vergleichsprüfung Mathematik 2009 – Rückmeldung der Ergebnisse", Wissenschaftliches Arbeitspapier des Instituts für Bildungsevaluation der Universität Zürich.

Formatiert: Deutsch (Schweiz)

CSS INSTITUT FÜR EMPIRISCHE GESUNDHEITSÖKONOMIE

Das „CSS Institut für empirische Gesundheitsökonomie“ ist eine Einrichtung der CSS Kranken-Versicherung AG, die der Forschung und Ausbildung dient.

Das Institut soll aufgrund von aktuellen und repräsentativen Datengrundlagen empirisch belegbare Antworten auf Fragen der effizienten Finanzierung und der gerechten Lastenverteilung von Gesundheitsleistungen liefern.

Die Forschungsergebnisse sind in geeigneter Art und Weise in die politische und wissenschaftliche Diskussion einzubringen.

Das Institut wurde Anfang 2007 von der Geschäftsleitung der CSS Kranken-Versicherung AG ins Leben gerufen.

Die Finanzierung erfolgt einerseits durch Mittel der CSS Kranken-Versicherung AG, andererseits und je nach Art des Forschungsprojekts durch Dritte.

Die wissenschaftliche Objektivität und Unabhängigkeit der Forschungstätigkeit misst sich an der Qualität und der Art der Publikationen und Präsentationen der Institutsmitarbeitenden.

Das Institut hat seinen Sitz in Luzern.