Bildungsprozesse quantitativ, objektiv und vergleichend darstellen

Kann man Bildung messen?

Schulleistungsstudien und Evaluationen pädagogischer Institutionen nehmen für sich in Anspruch, die Ergebnisse von Bildungsprozessen in Form von quantitativen Aussagen objektiv und vergleichend darstellen zu können. Aber kann man Bildung tatsächlich messen? Die folgenden Überlegungen gehen der Frage nach und plädieren für eine kritisch abwägende Position.

Autor |
Prof. em. Dr. Walter Herzog,
Institut für Erziehungswissenschaft, Abteilung Pädagogische Psychologie, Universität Bern

walter.herzog@edu.unibe.ch

www.walterherzog.ch

Auf die Frage, ob man Bildung messen kann, gibt es keine schnelle Antwort. Denn zuvor müsste geklärt sein, was unter Bildung und Messung zu verstehen ist. Das aber ist im Handumdrehen erst recht nicht zu leisten. Nicht ganz zu Unrecht wird dem Bildungsbegriff nämlich vorgeworfen, ein Containerwort zu sein, das sich fast beliebig mit Inhalt füllen lässt (vgl. Lenzen 1997). Und was den Messbegriff anbelangt, so wird unsere Erfahrung mit physikalischen Maßeinheiten kaum genügen, um zu entscheiden, ob sich Bildung messen lässt. Wenn wir die Frage nach der Messbarkeit von Bildung daher trotzdem stellen, dann steht uns ein etwas längerer Weg bevor, an dessen Ziel wir trotzdem nicht damit rechnen dürfen, eine eindeutige Antwort gefunden zu haben.

Die Zeit als Analogie

Wie wäre es, wenn wir eine andere Frage gestellt hätten: Kann man Zeit messen? Vermutlich würden wir weniger zögern, eine schnelle Antwort zu geben. Denn jede Uhr, auf die wir schauen, beweist nicht nur, dass man Zeit messen kann, sondern auch, dass sie gemessen wird. Allerdings würden wir einräumen wollen, dass die gemessene Zeit das Phänomen Zeit nicht voll abdeckt. Unsere Uhren wissen nichts vom Flug der Zeit, vom Leiden an der Zeit, von der verlorenen Zeit oder von der Zeitlosigkeit. Sie wissen auch nichts von den Modalitäten der Zeit. Wir können noch so lange auf ein Zifferblatt blicken und werden trotzdem nie erfahren, dass die vergangene Zeit eine qualitativ andere Zeit ist als die zukünftige Zeit. Was eine Uhr misst, ist daher nicht die ganze Wahrheit. Sollte etwas Ähnliches auch auf die Bildung zutreffen?

Wenn sich Zeit messen lässt: Wie wird das gemacht? Auch auf diese Frage wären wir bereit, relativ rasch zu antworten. Wir messen die Zeit, indem wir einen wiederkehrenden natürlichen oder künstlichen Vorgang als Standard setzen, um einen anderen Vorgang damit zu vergleichen (vgl. Elias 1984). Ob Sonnenuhr, Wasseruhr, Sanduhr, Räderuhr, Quarzuhr oder Atomuhr – immer gibt es einen sich wiederholenden Geschehensablauf, der als Vergleichsmaßstab dient, um einen anderen Geschehensablauf zu messen. Die Einheit der Zeitmessung ist dann die Dauer des periodischen Ablaufs, die per Konvention als Sekunde, Minute, Stunde, Tag oder Jahr bestimmt werden kann.

Das aber heißt, dass uns die metrische Zeit nicht einfach gegeben ist. Uhren mögen uns die Zeit zwar anzeigen, aber sie tun es nicht, indem sie eine an sich seiende Zeit abbilden. In Wahrheit auferlegen sie unserem Zeitempfinden eine bestimmte Ordnung, die unter pragmatischen Bedingungen steht. Menschen messen die Zeit, um sich besser in der Zeit orientieren zu können. Das zeigt die Geschichte der Zeitmessung. Noch bis ins 19. Jahrhundert waren die Uhren auf die Bedürfnisse einer Agrargesellschaft abgestimmt. Bei einem hohen Grad an Selbstversorgung und schwach entwickelten Marktbeziehungen genügte es, die Uhren nach dem Sonnenstand auszurichten, der von Ort zu Ort variiert. Das änderte sich mit den wachsenden Verflechtungen zwischen den Menschen aufgrund der Urbanisierung und Industrialisierung der Gesell-

schaft, der Expansion der Märkte und des technologischen Fortschritts in fast allen Lebensbereichen. Insbesondere der Ausbau des Eisenbahnnetzes und die rasante Entwicklung der Telegraphie waren für die Vereinheitlichung der Zeitmessung von wesentlicher Bedeutung. Die aufkommenden Nationalstaaten beförderten die Entwicklung, da deren Verwaltungsapparate generell auf einheitliche Maßsysteme angewiesen waren. Die Standardisierung der wichtigsten Maßeinheiten bereitete den Boden für den Eintritt der Gesellschaft ins „Zeitalter der Vergleichung", dessen Charakteristikum darin liegt, dass es praktisch keine menschlichen Leistungen mehr gibt, die nicht einem Vergleich unterzogen würden (vgl. Nietzsche 1997, S. 464 f.).

Beurteilen und Vergleichen

Die Begriffe der Standardisierung und der Vergleichung führen uns zurück zur Bildung. Denn insofern die pragmatische Funktion eines Maßsystems darin liegt, die menschliche Kommunikation von lokalen Grenzen zu befreien und für überregionale Vergleiche zu öffnen, verdankt sich auch das Ansinnen, Bildung zu messen, im Wesentlichen dieser Zielsetzung. Verstanden als Leistungsstandards, legen Bildungsstandards fest, welche Lernleistungen von Schülerinnen und Schülern erwartet werden, und zwar auf nationaler, wenn nicht auf internationaler Ebene (vgl. Herzog 2013). Um zu überprüfen, ob die Bildungsstandards erreicht werden, braucht es Messinstrumente, die eine solche vergleichende Prüfung im nationalen beziehungsweise internationalen Rahmen ermöglichen. In Bezug auf die Tradition des Bildungsdenkens, die wir mit Namen wie Wilhelm von Humboldt oder Georg Wilhelm Hegel in Verbindung bringen, hat der Fokus auf die Ergebnisse von Bildungsprozessen zwar ein reduziertes Verständnis von Bildung zur Folge, jedoch stellt die Beurteilung von Schülerleistungen für ein öffentliches Bildungssystem nichts grundsätzlich Neues dar. Prüfungen gehören genauso zum Normalbetrieb einer Schule wie die Zertifizierung von Leistungen mittels Noten. Ist damit die Frage nach der Messbarkeit von Bildung aber nicht bereits beantwortet? Was anderes als ein Messinstrument stellt denn eine Notenskala dar? Doch so einfach, wie es plötzlich den Anschein macht, ist es nicht.

Messen mithilfe von Skalen

Fragen wir etwas genauer, was eine Messung ist. Nach einer in den Sozialwissenschaften weit verbreiteten Definition beruht eine Messung auf der Zuordnung von Zahlen zu Phänomenen entsprechend einer Regel. So können wir Menschen mit der Eigenschaft, männlich oder weiblich zu sein, Zahlen zuordnen – den Männern eine Eins, den Frauen eine Zwei. Damit haben wir eine Messung vollzogen, allerdings eine äußerst primitive, denn die Zahlen erlauben lediglich eine nominelle Zuordnung. Man spricht deshalb von einer „Nominalskala". Die Werte einer Nominalskala haben dieselbe Funktion wie die sprachliche Benennung des Phänomens. Die Zuordnung der Zahlen ist zudem willkürlich und ließe sich im genannten Fall auch genau umgekehrt vornehmen.

Schon gehaltvoller ist die Zuordnung von Zahlen, wenn wir eine Rangreihe bilden können – zum Beispiel die Reihenfolge, in der die Teilnehmenden an einem Hundertmeterlauf im Ziel eintreffen. Wenn wir als Zahlensystem die natürlichen Zahlen nehmen und diese beginnend mit der Zahl 1 vergeben, dann definieren die Zahlen eine Rangordnung. Sie legen fest, wer der Schnellste, Zweitschnellste, Drittschnellste et cetera ist. Aus der Ordnung lässt sich aber nicht schließen, wer der oder die Schnellste schlechthin ist, da wir nur die Reihenfolge beim Eintreffen am Ziel registriert, aber nicht auch die Laufzeit gestoppt haben. Eine „Rang-" beziehungsweise „Ordinalskala" erlaubt es lediglich, Phänomene zu vergleichen, sagt aber nichts über die Distanzen zwischen den Skalenpositionen aus.

Genauer als eine Ordinalskala misst daher eine „Intervallskala", deren Zahlen nicht nur geordnet sind, sondern für gleiche Abstände (Intervalle) stehen. Davon wird bei einer Notenskala in der Regel ausgegangen. Der Abstand zwischen einer 1 und einer 2 ist gleich groß wie derjenige zwischen einer 3 und einer 4. Folglich lässt sich mit Noten im eigentlichen Sinne rechnen. Sie lassen sich addieren und subtrahieren, multiplizieren und dividieren. Zudem lassen sich Mittelwerte (Notendurchschnitte) berechnen, was im Falle einer Ordinalskala unzulässig wäre. Auch wenn mit Noten gerechnet wird, würde trotzdem niemand behaupten, der Nullpunkt einer Notenskala habe eine klare Bedeutung. Die Null müsste dann nämlich inhaltlich definiert werden, zum Beispiel als absolutes Nichtwissen in einem Fach. Aber wie könnten wir so etwas feststellen?

Das Kriterium eines definierten Nullpunktes ist erst bei einer „Ratioskala“ erfüllt, wie die gängigen physikalischen Maßsysteme für Länge, Gewicht oder Uhrzeit zeigen. Damit ist eine Messung möglich, die von lokalen Bedingungen gänzlich unabhängig ist, was es zum Beispiel bei einem Hundertmeterrennen erlaubt festzustellen, ob ein Rekord gebrochen wurde. Dank des absoluten Nullpunkts sind bei einer „Verhältnisskala“ (wie sie auch genannt wird) auch die Proportionen zwischen den Zahlen aussagekräftig. Zwei Stunden sind doppelt so viel wie eine Stunde, zehn Meter fünfmal mehr als zwei Meter, hundert Gramm zehnmal weniger als ein Kilogramm, et cetera.

Bildung quantitativ erfassen

Wenn wir Messung als Zuordnung von Zahlen zu Merkmalen von Objekten gemäß einer Regel definieren, dann sehen wir, dass die Regeln mit steigendem Skalenniveau strenger werden. Allerdings fragt sich, ob wir tatsächlich bei allen vier Skalentypen von Messung sprechen wollen. Die Zuordnung von Zahlen ist zwar eine notwendige Voraussetzung für eine Messung, aber weder eine Nominalskala noch eine Ordinalskala ist auf Zahlen angewiesen. Wir könnten auch Buchstaben verwenden, um die Skalenpositionen zu bezeichnen. Erst für Intervall- und Ratioskalen sind (reelle) Zahlen unabdingbar. Insofern Zahlen Größen darstellen, das heißt das Ausmaß zum Ausdruck bringen, in dem etwas existiert, und wir den Begriff der Messung in der Regel mit Quantifizierung in Verbindung bringen, scheint es in der Tat angemessener zu sein, erst ab dem Niveau einer Intervallskala von Messung zu sprechen (vgl. Michell 2004).

Die Grenze zum Messbegriff beim Übergang zu einer Intervallskala zu legen, macht auch deshalb Sinn, weil es bei der Messung von Bildung genau darum geht, nämlich Bildung quantitativ zu erfassen, um Lernleistungen über lokale Grenzen hinaus auf nationaler oder internationaler Ebene zu vergleichen. Anders als eine Klassifikation oder eine Rangreihe, die von schwer kontrollierbaren Einflüssen der Messsituation abhängig sind, gelten quantitative Daten als unabhängig von den Umständen, unter denen sie erhoben werden. Dabei wird den Zensuren vorgeworfen, genau dies nicht zu leisten. Die Notenskala entspreche lediglich einer Ordinalskala, weil in die Leistungsbewertung lokale Kriterien wie die Zusammensetzung der Schulklasse, das Anspruchsniveau der Lehrkraft oder die unterschiedliche Ausnutzung der Skalenbreite einfließen würden. Erst quantitative Daten seien hinreichend verlässlich, um jenes „government by comparison“ (Martens/Niemann 2010) zu ermöglichen, das wir seit den ersten PISA-Erhebungen beobachten können.

Tests als Messinstrumente

Die Kritik an der Notengebung ist allerdings nicht neu. Bereits 1921 stellte Edward Thorndike die Praxis, Schülerleistungen mittels Noten zu bewerten, aufgrund ihrer Unzuverlässigkeit in Frage. Noten kämen einer bloßen Meinungsäußerung gleich und würden den Anforderungen an einen Messvorgang nicht gerecht. Seine Forderung lautete, Meinung sei durch Messung zu ersetzen – „to replace opinion by measurement“ (Thorndike 1921, S. 378).

Unter „Messung“ verstand Thorndike den Einsatz von Tests. Tatsächlich ist es bis auf den heutigen Tag nur in wenigen Bereichen der Psychologie gelungen, echte („fundamentale“) Messungen durchzuführen (vgl. Michell 2004). Dies, weil sich psychische Phänomene im Unterschied zu physischen Phänomenen nicht direkt beobachten lassen. Was der Wissenschaft sinnlich gegeben ist, sind immer nur Verhaltensweisen oder Produkte von Verhaltensweisen, die sie als Indikatoren nutzt, um das interessierende psychische Merkmal zu erschließen. Wenn Messung darauf beruht, dass Phänomenen Zahlen zugeordnet werden, dann erweist sich diese Zuordnung im Falle psychischer Phänomene oft als ambivalent. Nur eine Theorie, die uns sagt, was ein psychologisches Konstrukt wie Persönlichkeit, Gedächtnis, Intelligenz oder Kompetenz genau bedeutet, kann deren Messung eindeutig begründen. Da die Theorien jedoch im Allgemeinen fehlen, behilft man sich mit psychometrischen Tests.

Standardisierte Tests

Tests erfüllen die Ansprüche an eine fundamentale Messung nicht – die Rede ist daher von einer vereinbarten oder willkürlichen Messung. Jedoch werden sie so konstruiert, dass sie einem Messverfahren nahekommen. Die üblichen Testkriterien – wie Objektivität oder Reliabilität – sind Ersatzkriterien, die darauf ausgerich-

tet sind, die Durchführung und Auswertung eines Tests zu standardisieren, womit der Anspruch einer Messung, nämlich von äußeren Einflüssen frei zu sein, wenigstens ansatzweise eingelöst werden kann. Das Ziel der Vergleichbarkeit über lokale Grenzen hinaus kann dann auch mit einem standardisierten Test erreicht werden, wie das folgende Zitat von Cronbach (1970, S. 27) deutlich macht: „A standardized test is one in which the procedure, apparatus, and scoring have been fixed so that precisely the same testing procedures can be followed at different times and places".

Allerdings ist die Qualität eines psychometrischen Tests mit der Messqualität einer modernen Uhr kaum vergleichbar. Als problematisch erweist sich insbesondere das Kriterium der Validität (vgl. Koretz 2009). Während Objektivität und Reliabilität formale Kriterien bilden, um die Verlässlichkeit eines Messvorgangs zu beurteilen, ist das entscheidende Qualitätskriterium inhaltlicher Art und betrifft die Frage, ob der Test misst, was er zu messen vorgibt. Anders als bei einer fundamentalen Messung, bei der die Frage der Validität auf der Basis einer bewährten Theorie über den Forschungsgegenstand beantwortet werden kann, lassen sich Tests nur nachträglich validieren, indem die Testergebnisse mit einem Außenkriterium korreliert werden. So werden Intelligenztests in der Regel mit der Schulleistung korreliert. Aber ist die Schulleistung ein gültiger Indikator für Intelligenz? Und an welchem Außenkriterium wird ein Schulleistungstest validiert? Nicht ein theoretisch fundierter Messvorgang befindet über die inhaltliche Gültigkeit eines Tests, sondern dessen Korrelation mit einem Kriterium, das letztlich nach Gutdünken ausgewählt wird. Zwar könnte man einwenden, dass uns auch die Zeit nicht anschaulich gegeben ist, aber im Falle der Zeit haben wir genau das, was bei der Konstruktion von psychometrischen Tests in der Regel fehlt, nämlich eine gut bestätigte (physikalische) Theorie, die erlaubt, die Uhrzeit als fundamentalen Messvorgang zu begründen.

Messen, was sich messen lässt

Kann man Bildung messen? Ja, offensichtlich kann man Bildung messen. Wie im Falle der Zeit stellt sich aber die Frage, ob das, was wir messen, auch dem entspricht, was wir messen wollen. Bildung auf das Ergebnis von schulischen Lernprozessen zu reduzieren, ist zwar nicht falsch, blendet aber vieles aus, was in der Tradition bildungsphilosophischen Denkens unter „Bildung" verstanden wird. Hinzu kommt die Frage der Messqualität. Wenn schon für eine Uhr gilt, dass sie nur so gut sein kann, „wie die dem Messprozess zugrundeliegende Theorie" (Eigen 1984, S. 216), dann muss für einen Schulleistungstest erst recht gelten, dass er nicht besser sein kann als die Theorie, die ihm zugrunde liegt. Doch mehr als ein paar Bruchstücke zu einer Theorie schulischen Lernens und schulischer Leistung haben wir nicht. Wie im Falle der Intelligenz und anderer psychologischer Konstrukte übernehmen die Tests die Funktion einer Ersatztheorie. Das mathematische Modell, das der Entwicklung eines Leistungstests zugrunde liegt, gibt dem Messgegenstand jene Struktur, die wir ihm mangels einer empirisch abgesicherten Theorie nicht geben können. Man fühlt sich an das Bonmot von Edwin Boring erinnert, wonach Intelligenz ist, was durch den Intelligenztest gemessen wird – „Intelligence is what the test tests" (Boring 1923, S. 35). Auch auf die Bildung scheint zuzutreffen, dass es letztlich die Tests sind, die festlegen, was unter Bildung zu verstehen ist.

Allerdings liegt das Problem nicht schon darin, dass die Tests festlegen, was Bildung ist, denn dies tun wir auch, wenn wir sprachlich über Bildung kommunizieren. Zwar bleibt immer etwas ausgeblendet, wenn wir ein Phänomen messen, aber vollständig erfassen können wir es auch nicht, wenn wir es qualitativ beschreiben. Das Problem der Leistungstests ist daher nicht, dass sie den Bildungsbegriff verengen, sondern dass wir nicht wissen, wie weit die Bildung, die uns in Form von Testwerten vorliegt, der Bildung entspricht, über die wir uns in Worten verständigen. Messung bringt zwar mehr Präzision in unsere Kommunikation, da sie uns zwingt, genauer zu sagen, was wir meinen. Doch die Genauigkeit des Ausdrucks ist noch keine Garantie dafür, dass wir dem Ausdruck auch die richtige Bedeutung geben. Objektivität im Sinne von intersubjektiver Übereinstimmung ist nicht dasselbe wie Objektivität im Sinne von Gegenstandsadäquatheit. Tests können uns helfen, die erste Art von Objektivität zu verbessern, tragen aber nicht zwingend zur Verbesserung der zweiten bei. Der Zugewinn an kommunikativer Präzision, den ein psychometrischer Test ermöglicht, kann erkauft sein mit einer Einbuße an Gewissheit darüber, was der Test misst. Solange wir in diesem Dilemma gefangen sind, ist fraglich, wie weit eine Bildung, die sich messen lässt, es auch wert ist, gemessen zu werden.

Literatur |
Boring, E.G.: Intelligence as the Tests Test It. In: The New Republic 1923 (35), H. 6, S. 35–37
Cronbach, L.J.: Essentials of Psychological Testing. New York 1970
Eigen, M.: Evolution und Zeitlichkeit. In: Chr. Link (Hrsg.): Die Erfahrung der Zeit. Stuttgart 1984, S. 215–237
Elias, N.: Über die Zeit. Hrsg. von M. Schröter. Frankfurt 1984
Herzog, W.: Bildungsstandards – eine kritische Einführung. Stuttgart 2013
Koretz, D.: Measuring Up. What Educational Testing Really Tells Us. Cambridge 2009
Lenzen, D.: Lösen die Begriffe Selbstorganisation, Autopoiese und Emergenz den Bildungsbegriff ab? In: Zeitschrift für Pädagogik 1997 (43), S. 949–968
Martens, K./Niemann, D.: Governance by Compariosn. How Ratings & Rankings Impact National Policy-Making in Education. TransState Working Papers No. 139. Bremen 2010
Michell, J.: Measurement in Psychology. Critical History of a Methodological Concept. Cambridge 2004
Nietzsche, F.: Menschliches, Allzumenschliches. In: Werke in drei Bänden, Erster Band. Darmstadt 1997, S. 435–1008 (Original 1886)
Thorndike, E.L.: Measurement in Education. In: Teachers College Record 1921 (22), S. 371–379