Datenbasierte Unterrichtsentwicklung mit VERA: (Wie) kann das funktionieren?

Ingmar Hosenfeld, Michael Zimmer-Müller und Josef Strasser

Abstract
Zentrales Ziel von in der Kultusministerkonferenz verabredeten Vergleichsarbeiten ist die Unterrichtsentwicklung. Der Beitrag setzt sich theoretisch mit Hindernissen und Gelingensbedingungen auseinander und entwirft eine Perspektive für eine Erhöhung der Nützlichkeit für die Schulen, indem eine zeitlich und thematisch deutlich flexiblere, computerbasierte Durchführung implementiert wird.

1. Ziele datengestützter Unterrichtsentwicklung: Worum geht es?

Mittlerweile werden seit zwanzig Jahren regelmäßig psychometrische Tests in Schulen geschrieben, die in den schulischen Jahresablauf integriert werden bzw. integriert werden müssen und als Vergleichsarbeiten (VERA), Kompetenztests, oder KERMIT (KompetenzenERMITeln) bezeichnet werden.

Die Bundesländer haben sich hierzu in der Gesamtstrategie zum Bildungsmonitoring zuletzt im Jahr 2015 verabredet. Darin sind auch Verfahren zur Qualitätssicherung auf Schulebene beschrieben. Letztere enthalten Sprachstandsmessungen für unterschiedliche Altersgruppen und landesspezifische Leistungsvergleichsuntersuchungen, aber auch als bundesweit verabredete Aufgabe, die Durchführung von länderübergreifenden Tests wie Lernstandserhebungen oder Vergleichsarbeiten in verschiedenen Jahrgangsstufen. »Vergleichsarbeiten sind Teil eines Bündels von Maßnahmen, mit denen die Länder eine evidenzbasierte Qualitätsentwicklung und -sicherung auf Ebene der einzelnen Schule gewährleisten« (KMK, 2015, S. 13) wollen. Die mit der Durchführung von Vergleichsarbeiten verbundenen Ziele sind schnell benannt: Es geht um »Unterstützung der Unterrichts- und Schulentwicklung jeder einzelnen Schule durch eine an den Bildungsstandards orientierte Rückmeldung als Standortbestimmung mit Bezug zu den Landesergebnissen« (KMK, 2015, S. 13). Darüber hinaus sollen »Vergleichsarbeiten eine wichtige Vermittlungsfunktion für die Einführung der fachlichen und fachdidaktischen Konzepte der Bildungsstandards« (KMK, 2015, S. 13) übernehmen.

Hierfür werden je nach Bundesland und durchführender Einrichtung[1] verschiedene Formen von Rückmeldungen angeboten. Letztlich handelt es sich bei diesen unterschiedlichen Darstellungsweisen oder Aufbereitungen bislang um die als richtig oder falsch codierten Antworten der Schülerinnen und Schüler bzw. darauf aufbauende Skalierungen. Diese liegen in Form von verschieden gestalteten Darstellungen und Aggregationen als sogenannte Lösungshäufigkeiten und Kompetenzstufenzuordnungen bzw. -verteilungen vor, die fächer- und zum Teil domänenspezifisch an die Schulen rückgemeldet werden, so etwa auf Ebene der Schülerinnen und Schüler als einzelne Werte. Auf Klassenebene können Lösungshäufigkeiten einzelner Items und diese im Vergleich zur eigenen Schule oder dem Land betrachtet werden sowie die Verteilung der Kompetenzstufe innerhalb der Klasse im Vergleich zu den anderen Gruppen, teilweise auch zu Schulen mit einer ähnlichen sozialen Zusammensetzung (so genannter »fairer Vergleich«). Dies ist die (Daten-)Basis, auf der Schulen nun ihre eigene Unterrichtspraxis überdenken und ggf. ändern sollen.

2. Blick in die Praxis: Wie gut funktioniert datenbasierte Unterrichtsentwicklung?

Zu dieser Frage liegt inzwischen eine Vielzahl von empirischen Studien vor (beispielsweise Dedering, 2011; Posch, 2009; Maier & Kuper, 2012; Diemer, 2013; Wurster, Richter & Lenski, 2017; Zimmer-Müller & Hosenfeld, 2013) – Terhart fasst sie pointiert zusammen: »Die Quintessenz der Studien ist äußerst ernüchternd« (2015, S. 72).

Über die Gründe dafür, dass das Ziel der Unterrichtsentwicklung mit VERA nicht oder nur in einem geringen Ausmaß (also an wenigen Schulen) erreicht wird, ist mittlerweile Vieles bekannt[2]. Als Gründe benennen Lehrkräfte beispielsweise sehr oft, dass der mit dem Verfahren verbundene Aufwand zu groß sei. Zur Adressierung dieses Kritikpunktes werden derzeit die Lernstandserhebungen auf computerbasierte Testungen umgestellt, was sowohl die logistische Vorbereitung

1 Die Bundesländer beauftragen mit der Testadministration ihre Landesinstitute oder die Universitäten Jena und Kaiserslautern-Landau.

2 Wir nehmen an dieser Stelle keine Differenzierung hinsichtlich der Veridikalität von Aussagen der schulischen Akteure vor, da die subjektiven Wahrnehmungen ausschlaggebend für die individuelle Handlungsbereitschaft sind. Mögliche Diskrepanzen zwischen subjektiven Wahrnehmungen und objektivierten Maßen werden beispielsweise im Bereich der Kompetenzen von Lehrkräften bei der Interpretation von Daten diskutiert, wobei bislang ungeklärt ist, welches Kompetenzniveau für gelingenden Umgang mit den Lernstandserhebungen eigentlich erforderlich ist.

(Druck von Testheften) als auch den Auswertungsaufwand (automatisierte Auswertung geschlossener Aufgabenformate) erheblich reduziert. Diese (und weitere Maßnahmen der Aufwandsreduktion, z.B. die externe Auswertung der Schülerantworten) führen zwar zu einer Reduktion des wahrgenommenen Aufwandes, nicht jedoch zu deutlich gesteigerten Maßnahmen der Unterrichtsentwicklung. Es wird also offensichtlich, dass darüber hinaus weitere Barrieren bestehen müssen. Neben einer grundsätzlich ablehnenden Haltung den standardisierten Schulleistungsmessungen gegenüber bestehen unseres Erachtens zwei wesentliche Problembereiche:

Erstens: Fehlendes Wissen darüber, wie und was im Anschluss an die Vergleichsarbeiten in der Schule getan werden kann – im Regelfall verbunden mit der Wahrnehmung, keine ausreichende Unterstützung für mögliche Folgeschritte zu haben. Diese Problematik wird z.B. in folgendem Ausschnitt aus einem der im Rahmen des Projektes WeSU[3] geführten Interviews mit einer Schulleitung illustriert:

> »Wie ich es vorher auch schon mal gesagt habe wir haben in der Regel mindestens ein durchschnittliches Ergebnis oder überdurchschnittlich. Aber auch wenn wir mal ein schlechtes Ergebnis haben, gibt es kein von außen so geführtes Unterstützungsangebot. Zumindest ist mir nichts bekannt. Es wird, also wir haben das Gefühl oder uns drängt sich das Gefühl auf, man macht es, weil man es machen muss, hat aber keinerlei Konsequenz von zusätzlicher Unterstützung, zusätzlichen personellen Ressourcen, was ja momentan sowieso ein großes Problem ist und auch keine zusätzliche Unterstützung von materiellen Unterstützung, also zusätzliches Lernmaterial oder ähnliches. Also das ist mein Eindruck. Aber es kann natürlich auch sein, dass ich das einfach noch nicht entdeckt habe.« (Interview M15G010C)

Obwohl praktisch alle Bundesländer flankierende Beratungs- und Fortbildungsangebote bereit halten und es zu den eingesetzten Aufgaben umfangreiche didaktische Kommentierungen und exemplarische Ausführungen zur konkreten Weiterarbeit im Unterricht gibt,[4] werden diese verfügbaren Ressourcen von vielen Lehrkräften entweder gar nicht oder als nicht hilfreich wahrgenommen. Dies verweist auf die nach wie vor bestehende Notwendigkeit, die vorhandenen Angebote besser

3 Das Projekt »Impulse für die Weiterentwicklung von Schule und Unterricht durch die Arbeit mit Ergebnissen aus Vergleichsarbeiten in Grund- und Sekundarschulen (VERA3 und VERA8)« (kurz WeSU) ist eine vom BMBF geförderte Interviewstudie, in der Lehrkräfte und Schulleitungen zum schulinternen Umgang mit den Lernstandserhebungen befragt werden (Förderkennzeichen 01JS2201).

4 Siehe hierzu die umfangreichen Materialien zu den einzelnen Testbereichen auf den Internetseiten des IQB unter https://www.iqb.hu-berlin.de/vera/aufgaben/ sowie die landesspezifischen Materialien unter https://www.iqb.hu-berlin.de/vera/Materialien/ (zuletzt abgerufen 23.06.2023), außerdem Bachinger, Krelle und Engelbert-Kocher (2022).

sichtbar zu machen. Wir gehen allerdings davon aus, dass hier nicht »mehr« Kommunikation, sondern eher eine stärker fokussierte Kommunikation zielführend ist. Die didaktischen Materialien werden beispielsweise leicht übersehen, weil sie in der Vielzahl der Download-Angebote übersehen werden.

Zweitens: Fehlendes Wissen darüber, wie (Ergebnisse aus) Vergleichsarbeiten genutzt werden können. Ein Argument, welches seit den ersten Tagen der Vergleichsarbeiten von Lehrkräften benannt wird, lautet, dass der Ertrag der Lernstandserhebungen zu gering ist. Hierunter lassen sich zwei Kernargumente fassen:

1. Der diagnostische Nutzen sei gering, weil die Lernstandserhebung die gleiche Rangfolge der Schülerinnen und Schüler reproduziere, die den Lehrpersonen auch aus den sonstigen schulischen Leistungsmessungen bekannt sei.
2. Der Ertrag sei gering, weil der Test nicht zum (implementierten) Curriculum passe.

Zur Illustration dieses Argumentes folgt ein weiterer Auszug aus demselben im Rahmen des WeSU-Projektes geführten Interview:

> »Das Problem ist, dass die Bildungspläne ja nicht bundeseinheitlich sind, VERA aber zum Teil bundeseinheitlich verfasst wurde. Ich bin nicht der Deutschmensch, ich bin der Mathematiker. Es kommen einfach Aufgaben vor, die im zeitlichen Ablauf noch gar nicht da sind, die einfach später gemacht werden. Daher ist (…) die Voraussetzung, dass VERA so wirken könnte, wie es sollte, wären einheitliche Bildungspläne für alle, immer zeitliche Vorgabe, dass ich sagen kann, wenn VERA stattfindet, müssten die Kinder nach Plan dieses und jenes Thema bearbeitet haben. Es kommen immer wieder Themen vor, wo bei uns im hauseigenen Curriculum einfach später vorgesehen sind. Und dann hat das natürlich keine Aussagekraft, weil die Kinder das noch nicht gemacht haben. Oder wir setzen einen anderen Schwerpunkt. Auch die Möglichkeit haben wir. Ich denke da immer nur in der Grundschule an die Wahrscheinlichkeitsrechnung, die ich persönlich nicht mag. Und ich weiß von den Kollegen auch, dass das kein Lieblingsthema ist. Dann wird es sowieso an den Rand geschoben, oft auch zum Schuljahresende noch zum Ausklang und das ist dann natürlich in VERA noch nicht abgedeckt. Wird aber in VERA relativ regelmäßig abgefragt. Also da sehe ich ein Problem, dass hier die Deckung von der Abfrage zum Bildungsplan nicht gewährleistet ist.« (Interview M15G010C)

Spätestens seit Ingenkamp (1969) ist aus Schulleistungsstudien bekannt, dass Notengebung und Testergebnisse innerhalb von Klassen hoch korrelieren; das Argument der Lehrkräfte ist also durchaus zutreffend. Ausgeblendet wird damit aber, dass für ein Bildungssystem, das die Zugänge zu weiteren Bildungsangeboten an Zertifikate der abgebenden Institutionen koppelt, dieser klassenzentrierte

Fokus deutlich zu kurz greift, um Gerechtigkeit, Fairness oder Chancengleichheit zu erzielen. Ingenkamp (1969) hat dafür den Begriff des »klasseninternen Bezugssystems« geprägt (siehe auch Baumert & Watermann 2000). Darüber hinaus wird in dieser Sichtweise das inhaltliche Potenzial der kompetenzorientierten Testung verschenkt: der genaue Blick auf das, was die Schülerinnen und Schüler tatsächlich gekonnt haben und was noch nicht, und inwieweit die in den Bildungsstandards festgelegten jeweiligen Kompetenzen erreicht werden oder nicht. Hierzu bieten die Begleitmaterialien des IQB umfangreiche Hinweise, d.h. die Rückmeldung beschränkt sich nicht nur auf die Frage, »Wo stehen meine Schülerinnen und Schüler?«, sondern liefert auch Ansätze zum »Was könnte der nächste Schritt sein?« (im Sinne des »feed forward« bei Hattie & Timperley, 2007). Die konsequente Kompetenzorientierung der verwendeten Testaufgaben erlaubt vielfältige Perspektiven auf die demonstrierten Leistungen, setzt allerdings fachdidaktische Expertise voraus.

Die oben zitierte Aussage im Interview weist auch darauf hin, dass in Schulen nicht alle gemäß der Bildungsstandards zu vermittelnden Kompetenzbereiche (hier konkret die mathematische Leitidee Daten, Häufigkeit und Wahrscheinlichkeit) bis zum Testzeitpunkt unterrichtet werden; es wird also ein Spannungsfeld zwischen intendiertem Curriculum (hier formuliert in den Bildungsstandards und Lehrplänen), dem implementierten Curriculum (was zu welchem Zeitpunkt und in welchem Umfang im Unterricht thematisiert wird) und dem realisierten Curriculum (über welche Kompetenzen die Schülerinnen und Schüler tatsächlich verfügen) deutlich. Die Differenz zwischen intendiertem und implementiertem Curriculum ist in den Vergleichsarbeiten systeminhärent, denn die Bildungsstandards formulieren Kompetenzen, die ein oder sogar erst zwei Jahre nach der Lernstandserhebung erwartet werden. Die Lehrkräfte stehen also vor einer schwierigen Interpolationsaufgabe, weil sie die Differenzen zwischen gemessenem Leistungsstand (realisiertem Curriculum) und in den Standards formuliertem Anspruch vor dem Hintergrund der Differenz zum tatsächlich erteilten Unterricht bewerten müssen, um zielführende Schlussfolgerungen für den künftigen Unterricht ziehen zu können.

3. Wie Daten hilfreich genutzt werden könnten: Ausblick

Vor diesem Hintergrund halten wir es für notwendig, die Aufgabestellung von Vergleichsarbeiten zu schärfen und die Durchführung zukünftig so zu gestalten, dass Schulen das Instrument an ihre Bedürfnisse und Erkenntnisinteressen anpassen können. Der bereits eingeschlagene Weg[5] der Umsetzung der Vergleichsarbeiten als

5 Das Team von Kompetenztest.de der Universität Jena und das Zentrum für Empirische Pädagogische Forschung (zepf) der RPTU Kaiserslautern – Landau für das Land Niedersachsen

computerbasierte Tests sollte weiter vorangetrieben werden. Dabei geht es nicht nur um die eigentliche Testdurchführung, sondern auch um eine möglichst weitgehend automatisierte Auswertung. Zu diesem Weg gehören konsequenterweise auch neue Formen von Rückmeldungen auf Grundlage einer erweiterten Datenbasis (z.B. Bearbeitungszeit, häufige Fehler und ähnliches), die weiteres fachdidaktisches Potenzial besitzen.

Während die Umsetzung der computerbasierten Testung bereits als politischer Wille in den Auftrag an das IQB ergangen ist, setzen weitere Elemente der skizzierten Weiterentwicklung den politischen Willen zum Umdenken voraus. Es gilt, das primäre Ziel der Vergleichsarbeiten (Impulse für die Unterrichtsentwicklung) als alleiniges zu akzeptieren und den Gedanken, auf Landesebene mit den Vergleichsarbeiten auch Bildungsmonitoring zu betreiben, endgültig zu verwerfen. Die zentrale Festlegung von Testzeitpunkt und Testinhalt ist Ausdruck dieses Monitoringgedankens, steht aber, wie oben dargestellt, im Widerspruch zum wahrgenommenen Nutzen auf Seiten der Schulen. Um diesen zu steigern, brauchen Schulen unseres Erachtens deutlich höhere Freiheitsgrade in der Ausgestaltung als dies aktuell der Fall ist. Hervorzuheben sind hier insbesondere die Fragen nach dem Wann, Wer und Was:

- Wann: Schulen brauchen Werkzeuge, die sie dann einsetzen können, wenn der Bedarf dazu offenkundig wird. Das bedeutet, dass entsprechende Kompetenztestungen nicht auf ein vorgegebenes Zeitfenster (oder bestimmte Klassenstufen) beschränkt sein sollten, sondern jederzeit durchgeführt werden können – dies eröffnet automatisch auch die Option, dieselben Schülerinnen und Schüler zu mehreren Zeitpunkten zu untersuchen und so Lernzuwächse in den Blick zu nehmen.
- Wer: Zum einen bezieht sich das auf die zu untersuchende Klassenstufe (datengestützte Unterrichtsentwicklung sollte auch jenseits der Klassenstufen 3 und 8 stattfinden), zum anderen aber auch auf die Frage, ob nicht ein inhaltlicher Fokus auf spezifische Subgruppen (besonders Leistungsstarke, besonders Leistungsschwache, Schülerinnen und Schüler mit nicht-deutschem Sprachhintergrund etc.) gerichtet werden soll, indem etwa nur Teile der Klassen getestet oder betrachtet werden. Ein solches Vorgehen erscheint vor allem dann angezeigt, wenn spezifische Maßnahmen der Unterrichtsentwicklung die entsprechende Subgruppe fokussieren und die gewonnenen Daten auch zur (formativen) Evaluation der Entwicklungsmaßnahme verwendet werden können. Darüber hin-

haben bereits 2016 mit der Umsetzung computerbasierter Tests begonnen, aktuell erfolgt die Vorbereitung der Umstellung auf vollständiges computerbasiertes Testen am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) der Humboldt-Universität zu Berlin. Dies soll bis spätestens 2029 in allen Fächern und Domänen abgeschlossen sein.

aus kann die mit einer Fokussierung auf eine Subgruppe verbundene Einschränkung der diagnostischen Breite zugunsten der diagnostischen Tiefe aber auch zu einer Verbesserung der Anschlussförderung der einzelnen Schülerinnen und Schüler führen, wenn hier besonderer Handlungsbedarf besteht.
- Was: Was genau wird untersucht? Welches Fach bzw. welche Domäne soll getestet werden? Welche spezifischen Inhalte oder Kompetenzen? Sollte die Vergleichsarbeit breit angelegt im Sinne eines Screenings sein, wie das bisher der Fall war, oder fokussiert auf spezielle Inhalte, dafür in großer Tiefe (siehe oben)? Schulen könnten somit eigene inhaltliche Schwerpunktsetzungen während des Schuljahres wählen und diese mithilfe von VERA evaluieren (siehe oben). Das Problem der Wahrnehmung einer mangelnden Passung zwischen implementiertem Curriculum und Test kann so erheblich reduziert werden.

Eine solche Flexibilisierung setzt allerdings auch einiges voraus: Auf Seiten der Schulen muss eine Verständigung zu allen Facetten dieser Freiheitsgrade stattfinden, denn die Prämisse, dass die Teilnahme verpflichtend ist, sollte unseres Erachtens nicht aufgegeben werden. Die notwendige gemeinschaftliche Formulierung von Erkenntnisinteressen, Schwerpunkten und Zielen stellt wichtige Elemente dessen dar, was Rolff nachhaltige Unterrichtsentwicklung nennt: Diese »basiert auf organisationalem Lernen; die Lehrkräfte einer Schule müssen sich über ihre Vorstellungen von Unterricht verständigen, die für ihre Realisierung notwendigen Schritte vereinbaren und die Kriterien definieren, anhand derer sie den Erfolg ihrer gemeinsamen Anstrengungen messen« (2015, S. 30). Auf politischer Ebene erfordert dies den Mut zu einem konsequenten Umbau des Instruments sowie die Mittel, hinreichende Mengen von Aufgaben entwickeln, pilotieren und fachdidaktisch kommentieren zu lassen, um verschiedene inhaltliche Schwerpunktsetzungen zu ermöglichen.

Förderhinweis

Die diesem Artikel zugrunde liegende Interviewstudie, in der Lehrkräfte und Schulleitungen zum schulinternen Umgang mit den Lernstandserhebungen befragt werden, wurde im Rahmen des Projektes »Impulse für die Weiterentwicklung von Schule und Unterricht durch die Arbeit mit Ergebnissen aus Vergleichsarbeiten in Grund- und Sekundarschulen (VERA3 und VERA8)« mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01JS2201 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen.

Literatur

Bachinger, A., Krelle, M., & Engelbert-Kocher, M. (2022). Beispieltestaufgaben mit fachdidaktischer Kommentierung. In A. Bachinger, M., Krelle, M., Engelbert-Kocher et al. (Hg.), *Zuhörkompetenzen messen. Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe* (S. 73–103). Waxmann.

Baumert, J., & Watermann, R. (2000). Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In J. Baumert, W. Bos, & R. H. H. Lehmann (Hg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 317–372). Leske+Budrich.

Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen. *Unterrichtswissenschaft, 39*(1), 63–82.

Diemer, T. (2013). *Innerschulische Wirklichkeiten neuer Steuerung. Zur Nutzung zentraler Lernstandserhebungen*. Springer VS.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Ingenkamp, K. (1969). Sind Zensuren aus verschiedenen Klassen vergleichbar? *betrifft: erziehung, 2*(3), 11–14.

KMK (2015). Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. Beschluss der 350. Kultusministerkonferenz vom 11.06.2015*. https://www.kmk.org/fileadmin/Dateien/pdf/Themen/Schule/Qualitaetssicherung_Schulen/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf; Stand 02.08.2023.

Maier, U., & Kuper, H. (2012). Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen – Überblick zum Forschungsstand. *Die Deutsche Schule 104*, S. 88–99.

Posch, P. (2009). Zur schulpraktischen Nutzung von Daten: Konzepte, Strategien, Erfahrungen. *Die Deutsche Schule, 101*(2), 119–135.

Rolff, H.-G. (2015). Formate der Unterrichtsentwicklung und Rolle der Schulleitung. In H.-G. Rolff (Hg.), *Handbuch Unterrichtsentwicklung* (S. 12–32). Beltz.

Terhart, E. (2015). Theorie der Unterrichtsentwicklung: Inspektion einer Leerstelle. In H.-G. Rolff (Hg.), *Handbuch Unterrichtsentwicklung* (S. 62–76). Beltz.

Wurster, S., Richter, D., & Lenski, A. E. (2017). Datenbasierte Unterrichtsentwicklung und ihr Zusammenhang zur Schülerleistung. *Zeitschrift für Erziehungswissenschaft, 20*(4), 628–650.

Zimmer-Müller, M., & Hosenfeld, I. (2013). Zehn Jahre Vergleichsarbeiten: Eine Zwischenbilanz aus verschiedenen Perspektiven. *Themenheft der Zeitschrift »Empirische Pädagogik«* 4.