

## STANDARDISIERTE LEISTUNGSMESSUNGEN / TESTS

Positionspapier

**Standardisierte Tests sind keine Gesamtbeurteilung. Sie messen immer nur das, was zu einem bestimmten Zeitpunkt, mit den entsprechenden Aufgabenstellungen für einen bestimmten Zweck gemessen werden kann. Wenn Testergebnisse gleichzeitig formativ, prognostisch und selektiv genutzt werden sowie womöglich noch zur Beurteilung von Schulen und Lehrpersonen eingesetzt werden, werden sie unbrauchbar. Einmalig gemessene Testergebnisse dürfen keinen selektiven Charakter für einzelne Personen haben, weil dann zu viel von der Tagesform abhängt.**

Der LCH hat bereits mehrmals und frühzeitig in Positionspapieren und Stellungnahmen in den Medien auf die teilweise äusserst problematischen und gefährlichen Entwicklungen beim Einsatz von standardisierten Tests hingewiesen (vgl. u. a. Positionspapiere vom 16. Juni 2012, 28. April 2012 und 18. Juni 2011). Mit diesem Positionspapier fasst der LCH den Stand der Erkenntnisse zu den Chancen und Gefahren von standardisierten Leistungsmessungen zusammen und formuliert fünf Forderungen.

### Chancen von Tests

Tests werden zu unterschiedlichen Zwecke konzipiert. Sie können systemische Veränderungen und Fortschritte von Bildungssystem oder Schulen aufzeigen dienen so der Rechenschaftslegung (z. B. PISA, ÜGK). Auf Ebene Klasse können sie der persönlichen Standortbestimmung von Schülerinnen und Schülern dienen und die Beurteilung von Lehrpersonen besser abstützen, was die Bildungsgerechtigkeit erhöhen kann (z. B. Stellwerk, Checks). In den vergangenen zehn Jahren sind für alle möglichen Zwecke Tests konstruiert worden, die nun teilweise aus Kostengründen oder fehlendem Fachwissen multifunktional eingesetzt werden (z. B. zur Beurteilung von SchülerInnen, Klassen, Lehrpersonen, Schulen oder Systemen).

### Tests sind der Zwilling der Output-Steuerung

Die Einführung von geleiteten Schulen mit gleichzeitigem Aufbau einer ergebnis- und wirkungsorientierten politischen Steuerung war international einer der wichtigsten inhaltlichen Treiber für die Testung von Leistungen von Schulen und Bildungssystemen. Weitere Treiber sind wirtschaftliche Interessen, die Freude an neuen technologischen Möglichkeiten und die weit verbreitete Lust an Rankings und Wettbewerb. Internationale und nationale Agenturen vergleichen Länder, Regionen und Schulen, um die „besten“ Systeme ausfindig zu machen und die „Leuchttürme“ als Kopiervorlagen zu empfehlen. Neueste Forschungen und Erfahrungen zeigen aber, dass Leistungsdifferenzen zu einem guten Teil kulturell bedingt sind und neue schulische Systemelemente jeweils lokal adaptiert werden müssen. Politische Steuerung kann nicht nur mit Vergleichszahlen geschehen, sie benötigt auch Zukunftsvorstellungen, Gestaltungsmöglichkeiten und Ressourcen.

### Standardisierte Tests sind keine Gesamtbeurteilung

Zahlenbasierte Tests suggerieren mit Zahlennoten eine hohe Verlässlichkeit. Trotz grossen methodischen Fortschritten zeigen Tests aber immer nur Ausschnitte von Potentialen. Zudem sind die Ergebnisse von der Tagesform abhängig. Die Versuchung ist gross, auch nur einmalig durchgeführte standardisierte Leistungsmessungen scheinbar chancengerecht für die Selektion von passenden Bildungsangeboten und als rekursifste Belege für Zeugnisnoten zu nutzen. Werden einmalig durchgeführte Tests durch das Beilegen in Zeugnissen und die nachfolgende Nutzung bei Lehrstellenbewerbungen überbewertet, verfehlen sie ihr Ziel. Weil Tests immer nur das messen, was gemessen werden kann, wird das Postulat einer umfassenden Menschenbildung durch eine breite Grundausbildung massiv eingeschränkt. Einzelne Menschen werden auf standardisiert messbare Leistungen reduziert. Ergebnisse aus einmalig durchgeführten Tests können wegen der üblicherweise schwankenden Tagesform zu Fehlschlüssen führen.

Für die individuelle Nutzung bei Übertritten, Lehrstellenbewerbungen oder als Zeugnisbeilage sollen Testungen deshalb zwei- bis dreimal durchgeführt werden. Zu einer validen summativen und prognostischen Gesamtbeurteilung gehören immer auch Daten und Beobachtungen über längere Zeiträume, aus verschiedenen Perspektiven und zu mehreren Dimensionen der Bildungsidee. Zeugnisse und Einschätzungen von Lehrpersonen sind für eine faire Gesamteinschätzung daher unabdinglich.

#### **Tests sollen gemäss ihrem Zweck genutzt werden**

Werden Tests zu summativen, formativen, prognostischen und selektiven Zwecken oder auf der systemischen und individuellen Ebene mehrfach genutzt, sind die Ergebnisse unbrauchbar. Leistungsmessungen für das Systemmonitoring und die Forschung sollen nicht gleichzeitig auch zur Beurteilung von einzelnen Schülerinnen und Schülern, von Lehrpersonen oder von Schulen genutzt werden. Die nicht zu vermeidenden widersprüchlichen Anreize und Interessenskonflikte verführen u. a. zu teaching and learning to the test oder zu beschönigenden Trickereien. Die Ergebnisse werden dann verfälscht und damit unbrauchbar.

#### **PISA soll Langzeitvergleiche ermöglichen**

Zusammen mit der EDK kritisiert der LCH die Veränderungen bei den PISA-Messungen 2015, welche einen Langzeitvergleich der PISA-Daten verunmöglichen. Die bisherigen Investitionen verlieren ihren Wert. PISA soll nicht zu einem beliebigen Länderranking werden, sondern zur Verbesserung der nationalen Bildungssysteme verlässliche Daten liefern und Langzeitvergleiche ermöglichen.

#### **Keine Datenberge ohne Ressourcen für Verbesserungsmaßnahmen**

Der LCH steht bisher hinter den Messungen zur Überprüfung der nationalen Bildungsziele ÜGK. Sollten die Kantone aber weiterhin in der Bildung abbauen und Budgets reduzieren, müssen diese Ausgaben in neuem Licht gesehen werden. Steuerungsdaten nützen nur etwas, wenn dann auch Massnahmen umgesetzt werden können. Mit knappen Finanzen wird dies nicht der Fall sein. Das Geld würde dann besser in das Lernen von Schulen untereinander investiert als in Datenberge ohne Wirkung.

#### **Datenschutz und Öffentlichkeitsprinzip stehen im Konflikt**

Für die ÜGK hat die EDK eine Datenschutzregelung erlassen, welche Rankings in Kantonen mit Vollerhebungen verhindern soll. Ergebnisse aus kantonalen Leistungstests (Checks Nordwestschweiz) wurden zwar ebenfalls in Gesetzen und Verordnungen vor Veröffentlichungen geschützt. Diese rechtlichen Sicherungen konkurrieren aber mit dem Öffentlichkeitsprinzip, das gemäss einem Rechtsgutachten des LCH mehr Gewicht hat als die kantonalen Schutzbestimmungen. Erst ein Bundesgerichtsurteil wird diese Frage definitiv klären. Unklar ist auch, ob kantonale Parlamente die Offenlegung nicht eines Tages erzwingen werden, indem sie die Reglemente oder Gesetze ändern.

#### **Forderungen des LCH**

1. Ergebnisse aus standardisierten Leistungstests werden entweder für summative (auch selektive und prognostische) oder für formative Zwecke und nicht gleichzeitig für mehrere Zwecke genutzt.
2. Ergebnisse aus Tests für das Systemmonitoring werden anonymisiert und nur für systemrelevante Zwecke und nicht für Schulrankings oder die Beurteilung von Mitarbeitenden genutzt.
3. Ergebnisse aus standardisierten Tests, die den Schülerinnen und Schülern mitgegeben werden, müssen zwei- bis dreimal durchgeführt worden sein, damit die Resultate nicht allein von der Tagesform abhängen.
4. Die Kantone sorgen dafür, dass die versprochenen Instrumente für die individuelle Förderung und Lernstandserhebung rasch vorliegen.
5. Nach Datenerhebungen müssen auch Verbesserungsmaßnahmen budgetiert werden. Der Aufwand darf nicht „kostenneutral“ zu Lasten des Schulbetriebs finanziert werden

**ANHANG**

Übersichten zu Einsatz und Nutzung von Tests

**A. Wo und wozu wird evaluiert?**

Problematisch sind Mehrfachnutzungen auf der horizontalen Ebene ("wozu")

ZWECK (wozu) REICHWEITE (wo)	summativ, z. B. Beeinflussung der Zeugnisnote, Abgabe der Ergebnisse an SchülerInnen, die dann von Lehrfirmen eingefordert werden	prognostisch / selektiv / promotionswirksam z. B. Lehrstellenbewerbungen, Gymiübertritte, Bewertung von Lehrpersonen	formativ / fördernd / Selbstevaluation im Unterricht eingesetzt
lokal / individuell	Stellwerk, Checks	Stellwerk, Checks BRNW weitere Tests (z. B. Multicheck, Basiccheck), Unternehmenstests	Stellwerk, Checks BRNW Aufgabensammlung
regional / kantonal	Checks BRNW*, Stellwerk*	Checks*, Stellwerk*	Checks, Stellwerk
national	ÜGK		Nachuntersuchungen ÜGK
international	PISA		Nachuntersuchungen PISA

\*Die Intentionen der Hersteller stimmen nicht immer mit der tatsächlichen Nutzung in Kantonen und Schulen überein  
Vgl. Positionspapier vom 28. April 2012 und 18. Juni 2011

**B. Wer und was wird evaluiert?**

Problematisch sind mehrfache Nutzungen von Tests auf der vertikalen Ebene („wer“)

BEZUGSNORM (was) GETESTETE (wer)	Normative bzw. kriteriumsorientierte Bezugsnorm (Vergleich mit Vorgaben)	Individuelle Bezugsnorm, (Fortschritte, Vergleich mit sich selbst)	Soziale Bezugsnorm, (Wettbewerb, Vergleich mit anderen)
Lernende	Checks, Stellwerk, Multicheck, Basiccheck, etc.	Checks, Stellwerk	Checks, Stellwerk, private Tests
Lehrperson	Checks, Stellwerk	Checks, Stellwerk	Checks, Stellwerk
Schule	Checks, Stellwerk	Checks, Stellwerk	Checks, Stellwerk
Gesamtsystem	ÜGK	ÜGK	ÜGK, Checks, Stellwerk

### C. Kritik am Einsatz von ausgewählten Tests

#### 1. PISA

Seit dem Jahr 2000 wird PISA von der OECD mit 15-jährigen alle drei Jahre für die Fächer M, NW, und Lesen durchgeführt. Durch Rankings von Ländern und Regionen soll ein vergleichender Wettbewerb ausgelöst werden, der zu Interesse an besonders erfolgreichen Systemen und zu Verbesserungen der eigenen Systeme führen soll.

LCH und EDK kritisieren die dem Testbetrieb geschuldete häufige Durchführung alle drei Jahre und die veränderten Durchführungsmodalitäten mit zeitweise nicht erklärten massiv veränderten Ergebnissen.

Mit den Abbaumassnahmen und Budgetrestriktionen im Bildungsbereich kommt es zu einem Ungleichgewicht von Datenerhebung und Verbesserungsmaßnahmen. Daten machen nur Sinn, wenn nachher auch die Mittel für Verbesserungsmaßnahmen zur Verfügung gestellt werden.

vgl. Stellungnahme LCH zu PISA vom 8. Februar 2017 und vom 6. Dezember 2016

PISA	+	-	+ / - ?
OECD international Stichprobe 15-jährige Länderranking internationale definierte Kriterien  Mathe, Naturwissenschaften, Leseverständnis	Systemvergleiche Zusatzerhebungen ICT, financial literacy  Stichprobe erlaubt keine Schulrankings	unklare oder nicht veränderbare Ursachen  internationale Kriterien ohne Bezug zu Nationalen Bildungszielen und Lehrplan 21  Wechsel der Testdurchführung (PC basiert, neue Skalierung)  Hinweise auf teaching to the test und Betrug (Hilfestellungen, Schülerauswahl)  15-jährige haben nicht überall gleich viele Schuljahre hinter sich	Länderranking Impakt? Wirkung? Druck auf internationale Standardisierung (Abschlüsse, Kompetenzbeschreibungen?)

## 2. ÜGK

Die Testung der ersten im Jahr 2011 in Kraft gesetzten nationalen Bildungsziele (Grundkompetenzen) ist angelaufen. Geplant ist eine regelmässige Überprüfung der nationalen Bildungsziele (ÜGK).

Der LCH befürwortet die ÜGK, so lange wie zugesichert Schulrankings verhindert werden und die Ergebnisse im System harmonisierende Wirkungen erzeugen. Kosten für Veränderungen und Systemanpassungen sollen in einem sinnvollen Verhältnis zu den Messungen stehen und dürfen nicht zu Lasten der Lehrpersonen gehen.

ÜGK	+	-	+ / - ?
EDK national Stichprobe In kleinen Kantonen Vollerhebung 6. und 9. Schuljahr (HarmoS 8 und 11)  Kantonsranking National definierte Kriterien (Bildungsziele HarmoS)  Mathe, Schulsprache (Orthografie, Leseverständnis), erste Fremdsprache	Systemmonitoring und Systemvergleiche auf der Basis der nationalen Bildungsziele (Kohärenz mit kriteriumsorientierter Bezugsnorm)  Stichprobe erlaubt keine Schulrankings, ausser in kleinen Kantonen  Ermöglicht ev. Anpassung der Ziele (Landessprachen)	unklare oder nur mit Kostenfolgen veränderbare Ursachen  Beschränkung der Bildungsidee auf messbare Hauptfächer  keine brauchbaren Rückmeldungen für beteiligte Schulen.	Kantonsranking Impakt? Wirkung?  Mediendiskussion?  Druck auf nationale Harmonisierung (Studentafeln, durchlässige Systeme)  Unklar, ab wann Ziele verändert werden können. Allenfalls lange Wartezeiten bis zur Testung der zweiten Fremdsprache nach dem 9. Schuljahr

## 3. Kantonale Testungen: Checks und Stellwerk

Die Tests werden von kantonalen Konsortien finanziert aber in den Kantonen unterschiedlich genutzt.

Checks, Stellwerk u. ä.	+	-	+/- ?
Bildungsregionen Übernahme durch weitere Kantone  tw. Kantonale Vollerhebungen meist vor Stufen- übergängen  Nutzung tw. für Bewertungen von Schulen, Lehr- personen und Schülern  Mathe, Schul- sprache	Aufgaben- sammlungen für formative Zwecke  Ergebnisse für formative Zwecke nutzbar  Standortbestim- mung und Ver- gleichsmöglich- keiten für Schulen und Lehrpersonen	Zweckvermischung bei der Nutzung in einzelnen Kantonen und Schulen: Systemebenen, Selbst- und Fremdevaluation, formativ / summativ / selektiv / prognostisch  fehlende Gerichtsurteile zum Vorrang des Öffentlichkeitsprinzips  Beschränkung auf messbare Hauptfächer Für formative Zwecke wenig geeignete Rückmeldungen für Schulen  Bei Nutzung für Zeugnisse, Übertritte und Bewerbungen erhält die Tagesform wegen der nur einmaligen Durchführung ein zu grosses Gewicht  tw. Zwang zur zeitgleichen Durchführung	Impakt? Wirkung?  Vermischung von Zielen vermindern Wirkung  Teaching & learning to the test?  Missbräuche? (z. B. für Mitarbeiter- beurteilung)