UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Robert Bosch Stiftung

Prüfungskulturen im Wandel

Ein internationaler Überblick zu Forschungsstand, Policy und Praxis

Prof. Dr. Anne Sliwka, Dominik Maxelon, Jule Eckert



Autor:innen
Prof. Dr. Anne Sliwka
Dominik Maxelon, Jule Eckert
Institut für Bildungswissenschaft, Universität Heidelberg

Förderung
Gefördert von der Robert Bosch Stiftung

Überblick
Im Zeitalter von gesellschaftlichen Disruptionen und Künstlicher Intelligenz geraten etablierte Lern- und Prüfungskulturen unter Legitimationsdruck. Der Bericht analysiert international vergleichend, wie sich Hochleistungsschulsysteme in diesem Zuge weiterentwickeln. Die Beispiele machen sichtbar, wie der Ausbau und die Messung komplexer Kompetenzen organisiert werden und wie die unterschiedlichen Systeme Spannungsverhältnisse, z. B. von Lernförderung und Selektionsfunktion, bearbeiten.

Hinweis zur Entstehung des Textes
Der vorliegende Text ist unter Einbezug von Impulsen aus dem Workshop „Prüfungskulturen im Wandel“ entstanden, der am 8. Mai 2026 in Berlin stattfand. Im Workshop wurden internationale Entwicklungen, bildungspolitische Handlungsspielräume und mögliche Perspektiven für die Weiterentwicklung schulischer Prüfungskulturen in Deutschland diskutiert. Das Autor:innen-Team bedankt sich für die wertvollen Beiträge, die aus diesem Workshop hervorgegangen sind.

An dem Workshop nahmen teil
- Jörg Droste
- Patricia Drewes
- Jule Eckert
- Carolin Genkinger
- Dr. Michaele Geweke
- Sandra Heidemann
- Dr. Julia Hugo
- Cornelia von Ilsemann
- Dr. Simon Kretz
- Dominik Maxelon
- Dr. Andreas Skouras
- Prof. Dr. Anne Sliwka
- Vincent Steinl
- David Tepaße
- Thorsten Wolk

Inhaltsverzeichnis

Vorwort 5

1 Einleitung 7

2 Die Komplexität von Kompetenzen 10
2.1 Systematisierung von Kompetenz und Kompetenz-Messverfahren 11
2.2 Vergleich zentraler Kompetenz-Frameworks 13

3 Kompetenzen prüfen 16
3.1 Was erwarten wir von Prüfungen 17
3.2 Grenzen traditioneller Prüfungsformate 19

4 Internationale Ansätze und Forschungsstand zur Leistungsbewertung komplexer Kompetenzen 21
4.1 Voraussetzungen für zeitgemäßes Prüfen 22
4.2 Kollektive Wirksamkeit von Lehrkräften *(Collective Teacher Efficacy)* 23
4.3 Authentische und performanzorientierte Prüfungen 23
4.4 Formatives Assessment und Lernprogressionen komplexer Kompetenzen 25
4.5 Leistungsbewertung in kooperativen Arbeitsformen 26
4.6 Bewertung kreativen Denkens und weiterer komplexer Kompetenzen 28

5 Internationale Praxisbeispiele zur Leistungsbewertung in Abschlussprüfungen 30
5.1 Finnland, Dänemark und Norwegen 31
5.2 Estland 34
5.3 USA 34
5.4 Kanada 35
5.5 Australien und Neuseeland 37
5.6 Singapur 38
5.7 International Baccalaureate (IB) 41

6 Potenziale von Künstlicher Intelligenz (KI) in der Prüfungskultur 46
6.1 Von computergestütztem Testing zu KI-basiertem Assessment 47
6.2 Grenzen und Validitätsprobleme KI-gestützter Bewertung 48
6.3 Systemische Implikationen für Prüfungskulturen 49

7 Handlungsperspektiven für Deutschland 51
7.1 Rechtliche Rahmen für formative und kombinierte Leistungsnachweise erweitern 52

7.2 Selektionsärmere Entwicklungsräume und curriculare Freiräume schaffen 53
7.3 Zusätzliche Kompetenznachweise erproben 53
7.4 Zertifikatskurs zu *Assessment Literacy* entwickeln 54

8 Fazit 55

Literatur 58

Impressum 66

Warum wir unsere Lern- und Prüfungskultur weiterentwickeln müssen

Unsere Gesellschaft und die Arbeitswelt befinden sich in einem tiefgreifenden Wandel, der auch vor unseren Schulen und der Lebenswelt von Kindern und Jugendlichen nicht haltgemacht hat. Die Digitalisierung, insbesondere die rasanten Fortschritte im Bereich der Künstlichen Intelligenz, verändern grundlegend den Umgang mit Wissen. Informationen sind jederzeit verfügbar, Routineaufgaben werden zunehmend automatisiert, und die Fähigkeit, Wissen kritisch einzuordnen, wird immer wichtiger. Gleichzeitig verstärken die fortschreitende Globalisierung und zunehmende gesellschaftliche Polarisierung die Komplexität und das Gefühl von Unsicherheit – bei Lernenden wie auch bei Lehrkräften.

Nicht zuletzt deshalb wird an vielen Stellen, unter anderem in der bereits vor zehn Jahren beschlossenen KMK-Strategie „Bildung in der digitalen Welt“, die Stärkung zukunftsrelevanter Kompetenzen eingefordert: kritisches Denken, Kreativität, Kollaboration und ein reflektierter Umgang mit Unsicherheit.

Als Robert Bosch Stiftung beschäftigt uns die Frage, wie Schulen Kinder und Jugendliche bestmöglich dabei unterstützen können, diese Kompetenzen zu entwickeln. Wir sind davon überzeugt, dass die Antwort auf diese Frage entscheidend von der Lern- und Prüfungskultur abhängt.

In unserer Arbeit begegnen wir zahlreichen Schulen, die neue Wege gehen. Sie fördern selbstorganisiertes Lernen, setzen auf Deeper Learning, stärken die Partizipation von Schülerinnen und Schülern und schaffen Lerngelegenheiten, die auf Eigenverantwortung, Zusammenarbeit und die Bearbeitung komplexer Fragestellungen ausgerichtet sind. Auch die Preisträgerschulen des Deutschen Schulpreises zeigen eindrucksvoll, wie solche Lernumgebungen gestaltet werden können.

Heutige Prüfungsformate, die in der Regel auf Einzelarbeit sowie auf die Wiedergabe von Wissen oder dessen Übertragung auf neue Kontexte ausgerichtet sind, stehen dabei manchmal im Widerspruch zu den Lernzielen von offeneren oder kooperativeren Lernformen. Sie erfassen jene komplexen Kompetenzen nur unzureichend, die für eine aktive Teilhabe an der Gesellschaft unerlässlich sind. Künstliche Intelligenz setzt heutige Prüfungsformate zusätzlich unter Druck – das gilt interessanterweise insbesondere für Formate wie Hausarbeiten oder Präsentationsprüfungen, die bislang eher als Bewertungsmöglichkeit komplexerer Kompetenzen galten. Zugleich eröffnet sie neue Möglichkeiten, Lernprozesse sichtbar zu machen, individuelles Feedback zu unterstützen und Prüfungen stärker auf die Anwendung und Reflexion von Wissen auszurichten.

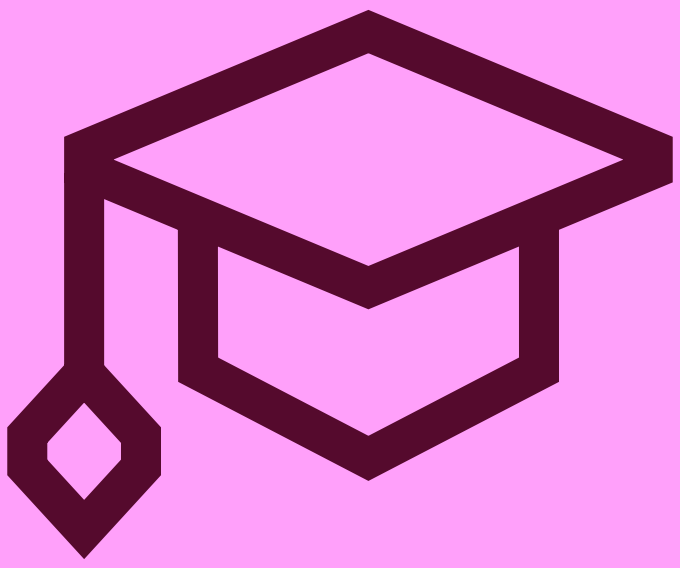
Die Diskussion über die Zukunft von Prüfungen ist deshalb weit mehr als eine Debatte über einzelne Formate. Sie berührt grundlegende Fragen darüber, welche Kompetenzen wir für bedeutsam halten und welches Lernen wir fördern möchten. Gleichzeitig erfüllen Prüfungen weitreichende gesellschaftliche Funktionen. Sie schaffen Orientierung, legitimieren Bildungsabschlüsse und eröffnen Zugänge zu weiteren Bildungs- und Berufswegen. Gerade deshalb ist die Weiterentwicklung von Prüfungskulturen keine rein pädagogische Frage, sondern eine gesellschaftliche Gestaltungsaufgabe.

Aus diesem Grund hat die Robert Bosch Stiftung die vorliegende Expertise bei Prof. Dr. Anne Sliwka und ihrem Team in Auftrag gegeben. Wir wollten verstehen, wie sich international führende und innovationsorientierte Bildungssysteme dieser Herausforderung stellen und welche konkreten Handlungsperspektiven sich daraus für Deutschland ergeben.

Die Expertise „Prüfungskulturen im Wandel“ liefert hierzu eine fundierte Analyse. Sie zeigt auf, dass eine Transformation der Prüfungskultur möglich ist, ohne auf anspruchsvolle Standards zu verzichten. Anhand von Beispielen aus Ländern wie Singapur, Finnland oder Kanada wird deutlich, wie neue, authentische Prüfungsformate, die Projektarbeit, Portfolios und kollaborative Leistungen einbeziehen, erfolgreich etabliert werden können. Dadurch wird sichtbar, wie die systemische Verankerung von komplexen Kompetenzen gelingen kann und wie die Spannung zwischen der pädagogischen Funktion des Förderns und Bewertens produktiv gestaltet wird. Darüber hinaus eröffnet sie einen differenzierten Blick auf die Chancen und Herausforderungen, die der Einsatz Künstlicher Intelligenz für die Weiterentwicklung von Leistungsbewertung mit sich bringt.

Wir sind überzeugt, dass Prüfungen dabei weit mehr als nur ein Instrument zur Leistungsmessung sein können – insbesondere, wenn sie formativ in den Lernprozess integriert sind. Wir hoffen, dass diese Publikation dazu beiträgt, die Debatte über die Zukunft der Leistungsbewertung in Deutschland zu bereichern. Sie richtet sich an Akteurinnen und Akteure aus Bildungspolitik, Bildungsadministration und Schulpraxis und soll dazu anregen, Prüfungen als das zu verstehen, was sie sein können: nicht nur Instrumente der Leistungsmessung, sondern wirksame Hebel für erfolgreiche Lernprozesse und gute Bildung.

Carolin Genkinger und Vincent Steinl
aus dem Team „Zukunftsfähiges Bildungssystem“
der Robert Bosch Stiftung



1 Einleitung

Gesellschaft, Arbeitswelt und globale Ordnungen verändern sich unter Bedingungen von Digitalisierung, Künstlicher Intelligenz (KI), Klimawandel sowie geopolitischen und medialen Machtverschiebungen tiefgreifend und rasant. Gleichzeitig finden sich in deutschen Schulen weiterhin Praktiken, die auf Vorstellungen von Lernen aus der Weimarer Republik zurückgehen (Reh et al. 2021). Aus diesen gehen Schüler:innen mit Fähigkeiten hervor, die vielfach nicht den Kompetenzen entsprechen, die sie in einer zunehmend komplexen und technologiegeprägten Welt benötigen werden.

Prüfungen entfalten erhebliche Rückwirkungen auf den Unterricht und strukturieren, was gelernt und gelehrt wird.

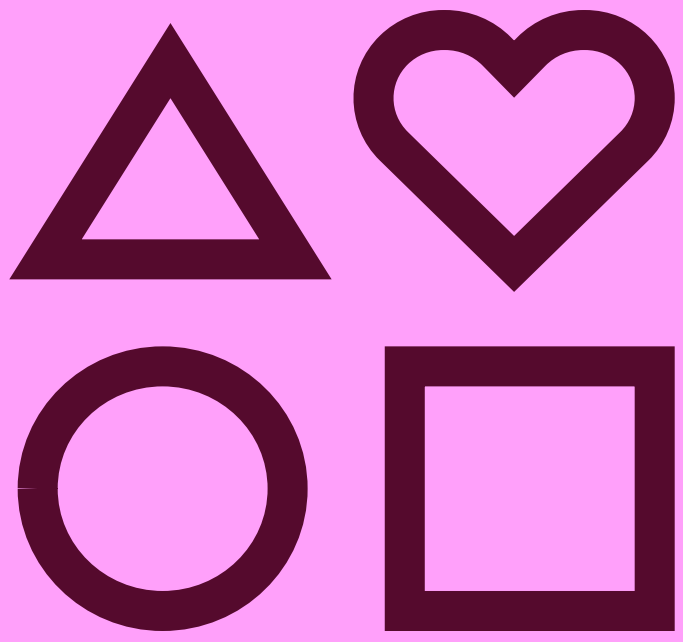
In Anbetracht dieser Paradoxie stehen Bildungssysteme vor der Aufgabe, Schüler:innen auf den kompetenten Umgang mit Unsicherheit und Komplexität sowie auf neue Formen menschlich-technologischer Zusammenarbeit vorzubereiten. Gerade deshalb gerät vor allem die aktuelle Prüfungskultur in Legitimationsnot: Denn Prüfungen entfalten erhebliche Rückwirkungen auf den Unterricht und strukturieren, was gelernt und gelehrt wird (Messick 1996). Sie entscheiden über Bildungswege und gesellschaftliche Zugänge, erfassen jedoch häufig nur jene Kompetenzen, die sich mit traditionellen Prüfungsformaten wie schriftlichen Klausuren vergleichsweise gut prüfen lassen (Darling-Hammond & Adamson 2010). Dabei wird zu selten die Frage aufgeworfen, ob wir mit diesen Formaten die Fähigkeiten erfassen können, mit denen Schüler:innen unsere Schulen verlassen sollen.

Im internationalen Vergleich zeigt sich, dass diese Frage intensiver bearbeitet wird und dass insbesondere leistungsstarke Bildungssysteme weltweit ihre Prüfungs- und Lernkulturen ebenso tiefgreifend transformieren wie ihre Gesellschaften. Dabei rücken neben fachlichem Wissen komplexe, überfachliche Kompetenzen, oft auch als 21st Century Skills oder Future Skills bezeichnet, stärker in den Fokus von Leistungsmessungen. Dazu zählen unter anderem kritisches Denken, Kreativität, Kommunikations- und Kollaborationsfähigkeit, komplexes Problemlösen und Selbstregulation (Pellegrino & Hilton 2012). Solche Kompetenzen werden insbesondere in Lernformaten gefördert, die mit realitätsnahen Problemstellungen, gemeinschaftlichen Arbeitsformen und authentischen Lernleistungen arbeiten.

Das deutsche Schulsystem kann aus Erfahrungen von Ländern wie Australien, Finnland oder Singapur lernen, um Schule zukunftsorientierter zu gestalten und Fehler zu vermeiden, aus denen dort bereits gelernt wurde. Dazu zählen Fragen, wie komplexe Kompetenzen systematisch gefördert werden können und wie den schulimmanenten Spannungen zwischen Lernförderung, Vergleichbarkeit, Fairness und Selektion angemessen begegnet werden kann.

Der vorliegende Bericht möchte diese Entwicklungen und Potenziale für den deutschsprachigen Raum sichtbarer machen und stellt daher

internationale Forschung, Policy-Entwicklungen und Praxisbeispiele zur Leistungsbewertung komplexer Kompetenzen vor. → Kapitel 2 klärt zunächst den Kompetenzbegriff und ordnet zentrale Kompetenz-Frameworks ein. → Kapitel 3 analysiert die Funktionen schulischer Prüfungen sowie die Grenzen traditioneller Prüfungsformate. → Kapitel 4 stellt internationale Forschungsstände zu zeitgemäßem Prüfen, formativer Leistungsbewertung, performanzorientierten Aufgaben, Kollaboration und Kreativität dar. → Kapitel 5 untersucht ausgewählte Praxisbeispiele aus leistungsstarken und innovationsorientierten Bildungssystemen, darunter Finnland, Dänemark, Norwegen, Estland, die USA, Kanada, Australien, Neuseeland, Singapur sowie das International Baccalaureate (IB). → Kapitel 6 diskutiert Potenziale und Grenzen KI-gestützter Assessmentverfahren. Anknüpfend formuliert → Kapitel 7 Handlungsperspektiven für Deutschland, bevor → Kapitel 8 die zentralen Spannungsverhältnisse und Entwicklungslinien zusammenführt.



2
Die Komplexität von Kompetenzen

2.1 Systematisierung von Kompetenz und Kompetenz-Messverfahren

Der Kompetenzbegriff fungiert als zentraler Referenzpunkt der Schulentwicklung, bleibt im bildungspolitischen, psychologischen und pädagogischen Diskurs jedoch vieldeutig. Erpenbeck und Kolleg:innen (2017) greifen diese begriffliche Pluralität auf, indem sie unterschiedliche Kompetenzverständnisse vergleichend systematisieren und deren theoretische Grundannahmen sowie Zielperspektiven herausarbeiten. Auf die begriffliche Unschärfe reagieren sie damit nicht durch eine normative Festlegung, sondern durch eine differenzierende Ordnung verschiedener theoretischer Zugänge.

Zentral ist die Unterscheidung von vier Definitionsclustern. Im ersten Cluster erscheinen Kompetenzen als ökonomisierte Variante des humanistischen Bildungsbegriffs; der Fokus verschiebt sich hier von Persönlichkeitsbildung hin zu Verwertbarkeit, *Employability* und outputorientierter Steuerung. Im zweiten Cluster wird Kompetenz sehr weit gefasst und nahezu synonym mit Handlungsfähigkeit verwendet: Kompetent ist, wer in einer Situation angemessen handeln kann. Das dritte Cluster definiert Kompetenzen als kontextspezifische, domänengebundene (also in bestimmten Fachdisziplinen relevante) kognitive Leistungsdispositionen. Dieses Verständnis prägt internationale *Large-Scale Assessments* wie PISA, TIMSS oder PIRLS maßgeblich. Kompetenzen werden hier funktional bestimmt und auf kognitive Problemlösefähigkeiten begrenzt; motivationale, affektive oder volitionale Dimensionen erfolgreichen Handelns bleiben ausdrücklich ausgeklammert.

> **Kompetenzentwicklung setzt an der individuellen Selbsttätigkeit der Lernenden an und entzieht sich folglich einer vollständigen Standardisierung.**

Das vierte Cluster schließlich versteht Kompetenzen als Selbstorganisationsdispositionen. In dieser systemisch-konstruktivistischen Perspektive sind Kompetenzen Fähigkeit und Bereitschaft zur eigenständigen Bewältigung komplexer und unsicherer Situationen. Kognitive, motivationale, emotionale und soziale Aspekte werden dabei integriert statt isoliert betrachtet. Arnold (2018) argumentiert im Sinne dieses Clusters, dass „kompetenzbildendes Lernen davon abhängig ist, ob es gelingt, die Vielfalt der Aneignungslogiken der Subjekte zunächst zu respektieren und ihre Aneignungsaktivitäten zu fördern“ (S. 119). Kompetenzentwicklung setzt damit an der individuellen Selbsttätigkeit der Lernenden an und entzieht sich folglich einer vollständigen Standardisierung. Während Kompetenz-

entwicklung auf die Förderung individueller, selbstorganisierter Lernprozesse zielt, orientieren sich Standardisierung und Inputsteuerung an normierten Vorgaben und steuerbaren Outputs. Beide Logiken markieren damit gegenläufige Pole.

Diese vier Cluster sind nicht als hierarchische Wertung zu verstehen, sondern als analytische Landkarte unterschiedlicher theoretischer Annahmen. Sie ermöglichen es, bestehende Kompetenzmodelle und Prüfungsformate systematisch zu verorten.

Vor diesem Hintergrund ordnen Erpenbeck und Kolleg:innen (2017) internationale Schulleistungsstudien eindeutig dem dritten Cluster zu. Auch neuere Erweiterungen, etwa das PISA-Assessment zu *Creative Thinking*, verbleiben in der Logik kognitiver Leistungsdefinitionen. Gemessen werden dort offene und zielgerichtete Denkstrategien, Problemlöseroutinen und fachspezifische Performanz unter standardisierten Bedingungen. Die Stärke dieses Ansatzes liegt in seiner hohen Vergleichbarkeit, Skalierbarkeit und teststatistischen Präzision. Zugleich weisen Erpenbeck und Kolleg:innen (ebd.) darauf hin, dass selbstorganisierte Handlungsfähigkeit, personale Dispositionen oder kreative Selbststeuerung in solchen primär kognitiven Assessments nur randständig oder gar nicht erfasst werden. Die Kritik richtet sich daher weniger gegen die methodische Qualität der Verfahren als gegen die Reichweite der aus ihnen abgeleiteten Interpretationen und Konsequenzen, die von Bildungsforscher:innen bereits kritisch eingeordnet wurden (Meyer & Zahedi 2014).

Erpenbeck und Kolleg:innen (2017) schlagen zudem eine kontextorientierte Neubestimmung der Anforderungen an Kompetenzmessung vor. Die klassischen Gütekriterien Objektivität, Reliabilität und Validität werden dabei nicht verworfen; es wird jedoch die Frage aufgeworfen, ob diese Kriterien bei allen Messverfahren gleichermaßen erreichbar und sinnvoll sind. Für komplexe, selbstorganisierte Handlungsdispositionen erweisen sich vollständig standardisierte, hoch objektive Verfahren nur begrenzt als angemessen. Stattdessen plädieren die Autor:innen für realistischere Anforderungen, die auch die Angemessenheit des Kontextes, die Transparenz der Bewertungsmaßstäbe, soziale Validität und die Anschlussfähigkeit an reale Handlungssituationen berücksichtigen. Der Einsatz unterschiedlicher Anforderungen an Messungen erscheint damit nicht nur als technisch-methodische, sondern zugleich als erkenntnistheoretische Frage.

Für die weitere Diskussion folgt daraus, dass Leistungsmessungen beziehungsweise Prüfungen unterschiedliche Zielsetzungen verfolgen können und entsprechend austariert werden müssen. Dies schließt die klassischen Gütekriterien ein, die in der Praxis nicht unabhängig voneinander maximiert werden können: Eine stärkere Standardisierung erhöht in der Regel Objektivität und Reliabilität, kann jedoch die Konstruktvalidität – also die inhaltliche Passung – bei komplexen, kontextgebundenen Leistungen einschränken. Umgekehrt geht eine ganzheitlichere Erfassung von Kompetenz meist mit geringerer Standardisierbarkeit einher. Die Überführung komplexer Kompetenzen in Prüfungsformate ist daher stets mit einem Spannungsfeld verbunden, dessen Konsequenzen reflektiert und je nach Funktion der Prüfung unterschiedlich gewichtet werden müssen.

2.2 Vergleich zentraler Kompetenz-Frameworks

Die disruptive Transformationsdynamik gesellschaftlicher, wirtschaftlicher und politischer Entwicklungen hat die Auseinandersetzung mit der Frage beschleunigt, welche überfachlichen Kompetenzen Schule heute fördern soll. Wir verwenden in diesem Zusammenhang den Begriff „komplexe Kompetenzen“ (University of Melbourne 2025), weil er im internationalen Diskurs vergleichsweise theoretisch offen ist. Er ermöglicht es, unterschiedliche Kompetenzkonzeptionen zu adressieren, ohne sich frühzeitig auf ein bestimmtes Framework oder eine spezifische Modelllogik festzulegen. Eine Übersicht konkurrierender Begrifflichkeiten findet sich bei Brandt und Kolleg:innen (2025). Wo internationale Praxisbeispiele mit anderen Terminologien arbeiten, greifen wir diese länderspezifischen Bezeichnungen auf.

In unterschiedlichen gesellschaftlichen Teilsystemen werden Kompetenzansätze verhandelt, die auf variierenden Problemwahrnehmungen sowie unterschiedlichen normativen Orientierungen und Zukunftsentwürfen beruhen. Je nach Perspektive ergänzen sich wirtschaftliche, humanistische und bildungspolitische Zielsetzungen – oder sie verschwimmen ineinander. Einen Überblick über nationale Frameworks bietet *FutureSkills4U* (Next-Education, o. J.), über internationale Frameworks *Harvard Education* (Harvard University, o. J.).

Was diese Frameworks verbindet, ist die Annahme, dass komplexe Kompetenzen ein Bündel kognitiver, metakognitiver, sozialer und emotionaler Fähigkeiten darstellen, das Lernende zur erfolgreichen Problemlösung in neuartigen Situationen befähigt (Binkley et al. 2012; Pellegrino & Hilton 2012). Im Unterschied zu Fachkompetenzen sind diese Fähigkeiten interdisziplinär angelegt; sie ermöglichen es, Wissen aus einzelnen Domänen transfer- und anwendungsfähig zu machen. Deshalb ist häufig auch von *Transferable Skills* oder Schlüsselkompetenzen die Rede. Zu den diskursprägenden Aufzählungen zählen gerade im deutschsprachigen Raum etwa die „4K“: kritisches Denken, Kreativität, Kommunikation und Kollaboration.

Die 4K gehen aus dem Vier-Dimensionen-Modell von Bildung des Center for Curriculum Redesign (CCR) hervor, das aus Wissen, Skills, Charakter und Meta-Lernen besteht und mit internationalen Kompetenzrahmen wie dem OECD *Learning Compass 2030* verknüpft ist (Bialik & Fadel 2017; OECD 2020). Die 4K sind in diesem Modell vor allem der Dimension der Skills zugeordnet. Bildung wird dabei als integriertes Zusammenspiel mehrerer Dimensionen verstanden: Fachliches Wissen wird nicht isoliert betrachtet, sondern in die Entwicklung überfachlicher Fähigkeiten, personaler Dispositionen sowie in die Fähigkeit zur Selbstregulation und reflexiven Steuerung des eigenen Lernens eingebettet.

Im Kontext einer „Bildung für das Zeitalter von KI“ (im Original: *Education for the Age of AI*) bleiben diese vier Dimensionen in ihrer Grundstruktur

erhalten, werden jedoch neu akzentuiert (Fadel et al. 2024). CCR unterscheidet zwischen Kompetenzbereichen, die angesichts disruptiver technologischer Entwicklungen für Menschen an Bedeutung gewinnen, und solchen Fähigkeiten, die voraussichtlich an KI-Systeme delegiert werden können. Im Bereich der Skills rückt mit Blick auf Kreativität insbesondere die menschliche Vorstellungskraft in den Vordergrund; im Bereich des kritischen Denkens wird fundierte Entscheidungsfähigkeit betont. Dialog erscheint als dezidiert menschliche Kommunikationsform, während im Feld der Kollaboration insbesondere *Leadership*-Fähigkeiten als Ressource für ko-konstruktives Handeln mit KI hervorgehoben werden. In der Dimension des Charakters werden vor allem Weltoffenheit, Mut sowie ethische Orientierungen wie Fairness und Verantwortungsbewusstsein als zukunftsrelevant hervorgehoben. Innerhalb der Dimension des Meta-Lernens gilt insbesondere die Lern- und Anpassungsfähigkeit als zentrales Unterscheidungsmerkmal zwischen Mensch und Künstlicher Intelligenz, das systematisch gefördert werden soll (ebd.).

Neben diesen international anschlussfähigen Modellen existieren weitere Frameworks, die den Diskurs um Zukunftskompetenzen für den deutschsprachigen Raum systematisieren. Die Allianz für *Future Skills* formuliert beispielsweise für das Jahr 2030 insgesamt 30 Kompetenzen, die künftig an Bedeutung gewinnen sollen (Rampelt et al. 2024; Rampelt et al. 2025). Das von Expert:innen aus Wirtschaft, Wissenschaft und Zivilgesellschaft entwickelte Framework unterscheidet grundlegende, transformative, gemeinschaftsorientierte, digitale und technologische Zukunftskompetenzen. Kritisches Denken und Kommunikation erscheinen auch hier als Querschnittsfähigkeiten. Hinzu kommen weitere Fähigkeiten wie Informationskompetenz, Selbstregulation, technologische Kompetenz sowie kulturelle und ethische Urteilsfähigkeit (Griffin & Care 2015; Pellegrino & Hilton 2012).

Im Bildungsansatz des *Deeper Learning*, der auf tiefes Verstehen und die Anwendung von Wissen zielt, stehen ebenfalls die 4K im Vordergrund. Schüler:innen sollen vernetztes Wissen aufbauen und lernen, kreativ zu denken, Probleme zu lösen und in unterschiedlichen Kontexten handlungsfähig zu werden, statt lediglich oberflächliches Faktenwissen zu reproduzieren (Pellegrino & Hilton 2012; Sliwka & Klopsch 2022). Eine Studie des National Research Council definierte *Deeper Learning* als die Verbindung solider Fachkenntnisse mit überfachlichen Kompetenzen, die es ermöglichen, Wissen flexibel in neuartigen und komplexen Problemlagen anzuwenden (Pellegrino & Hilton 2012). Diese Vision spiegelt sich in zahlreichen internationalen Bildungsrahmen wider, etwa im Kompetenzmodell der OECD – dem *OECD Lernkompass 2030* (OECD 2020) – oder in nationalen Curricula wie dem *21st Century Competencies*-Rahmen Singapurs (Tan et al. 2017) und den *Transversal Competences* in Finnland (Finnish National Agency for Education 2026a).

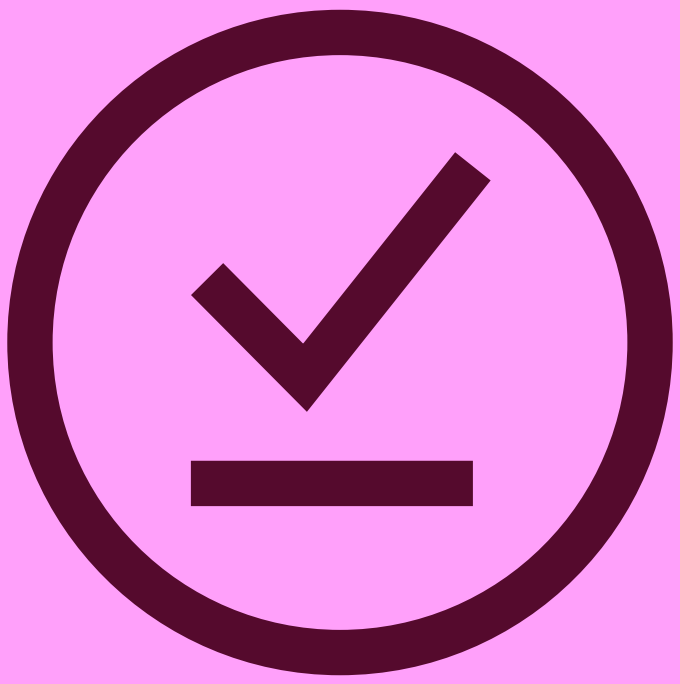
Komplexe Kompetenzen sind häufig mehrdimensional, nicht trennscharf und entwickeln sich dynamisch. Kritisches und kreatives Denken überlappen etwa, ebenso wie fachliches Wissen, kollaboratives Arbeiten und Problemlösen in authentischen Situationen meist ineinandergreifen. Forschung zeigt, dass der Erwerb solcher Kompetenzen deutlich weniger

linear verläuft als das Erlernen von Basiskompetenzen in Sprache und Mathematik, das sich oft entlang klarerer *Learning Progressions* organisieren lässt (Care et al. 2018; Shepard et al. 2013).

Trotz unterschiedlicher Akzentsetzungen besteht weitgehend Konsens darüber, komplexe Kompetenzen stärker ins Zentrum von institutionalisierten Bildungsprozessen zu rücken und fachliche Kompetenzen sowie deklaratives Fachwissen dann stärker als bisher als integrale Bestandteile übergeordneter Kompetenzentwicklungsziele zu begreifen.

> **Die Förderung komplexer Kompetenzen erfordert deutlich komplexere Prüfungsformate als jene, die im Schulsystem bislang dominieren.**

Als Hypothese dieses Papiers gilt daher: Die Förderung komplexer Kompetenzen erfordert deutlich komplexere Prüfungsformate als jene, die im Schulsystem bislang dominieren. Daraus ergibt sich die Frage, wie unterschiedliche Kompetenzverständnisse in konkrete Prüfungsformate übersetzt werden können.



3
Kompetenzen prüfen

3.1 Was erwarten wir von Prüfungen?

Schulisches Prüfen ist durch unterschiedliche Funktionen gekennzeichnet: Einerseits soll Leistungsbewertung Lern- und Entwicklungsprozesse unterstützen und fördern. In diesem Sinn erfüllt sie eine pädagogische Funktion, die im aktuellen Diskurs häufig als *Assessment for Learning* beschrieben wird. Andererseits hat Prüfen eine selektive Funktion, weil über Zensuren Bildungswege strukturiert und Zugänge zu weiterführenden Bildungs- und Berufsoptionen eröffnet oder verschlossen werden. Auch rechtliche Regelungen betonen zwar primär die pädagogische Zielsetzung schulischer Leistungsbewertung, zum Beispiel in Form von „Bestätigung, Ermutigung, Hilfe zur Selbsteinschätzung und Korrektur“ (Niedersächsisches Kultusministerium 2015). Dabei belegen empirische Studien die hohe Bedeutung von Feedback und Rückmeldungen für wirksame Lernprozesse (Hattie 2008; Helmke 2022). Ebenso zeigen zahlreiche Untersuchungen, dass die Praxis des Notengebens Lernprozesse hemmen, Motivation untergraben und lernförderliche Rückmeldeprozesse überlagern kann (Jürgens & Lissmann 2015; Rheinberg & Engeser 2018). Beutel (2024) weist darauf hin, dass es bei Reformen der Leistungsbeurteilung nicht um eine einfache Gegenüberstellung „harter“ und „weicher“ Formen geht, sondern um die Frage, welche Formen der Leistungsbewertung jeweils als sachlich angemessen und pädagogisch wirksam gelten können.

Im schulpädagogischen Reformdiskurs dominiert die Kritik an der allokativen Funktion von Schule und am Notensystem. Als „Märchen für Erwachsene“ bezeichnet Wampfler (2024b) das der traditionellen schulischen Leistungsbewertung zugrunde liegende „Primat der Meritokratie“ und die damit verbundene Perspektive auf individuelle Leistung ohne eine gleichzeitige Analyse und Bewertung des familiären, sozio-ökonomischen und gesellschaftlichen Kontextes, der diese Leistung möglich macht. Die Selektionsfunktion erhält ihre Legitimation aus der Vorstellung, Leistungen seien immer eindeutig individuell zurechenbar und vergleichbar (Wampfler 2024a). Schulische Leistung ist jedoch kein vorab vorhandenes individuelles Merkmal, sondern entsteht im Bewertungsprozess und wird sozial koproduziert. Selbst unter scheinbar isolierenden Prüfungsbedingungen bleiben Leistungen Ausdruck kollektiver Lern- und Unterstützungszusammenhänge von Peers, Lehrpersonen und Eltern; damit gerät die meritokratische Argumentationslinie unter Druck (Verheyen 2018).

Viele Konzepte lernförderlichen Feedbacks oder zeitgemäßen Prüfens blenden die Selektionsfunktion analytisch aus. Dass sie in der Theorie marginalisiert wird, bedeutet jedoch nicht, dass ihre Wirkungen in der Praxis verschwinden. Vielmehr zeigen Bialik und Fadel (2017), dass standardisierte Prüfungs- und Zulassungssysteme, etwa im Kontext des Hochschulzugangs, weiterhin zentrale Auswahlmechanismen darstellen, die Bildungsbiografien strukturieren. Solche Selektionssituationen wirken stark auf Curricula und Unterricht zurück, weil sie definieren, was als relevant gilt. Dadurch verengt sich schulische Praxis häufig auf vermeintlich

vergleichbare Prüfinhalte (*Teaching to the Test* beziehungsweise *Washback*-Effekt; Messick 1996). Zugleich verweisen Bialik und Fadel (2017) auf die Grenzen von Prüfungen: Leistungsbewertung ist historisch gewachsen, normativ aufgeladen und von Spannungen geprägt, insbesondere zwischen Selektion und Lernförderung. Diese Spannung wird durch bildungspolitische Steuerungslogiken weiter stabilisiert. So betonen sowohl das Bundesverfassungsgericht (2017) in seinem Urteil zur Studienplatzvergabe als auch die Beschlüsse der Kultusministerkonferenz (2024) zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung Anforderungen an Vergleichbarkeit, Standardisierung und rechtssichere Leistungsbewertung. Gütekriterien wie Objektivität, Reliabilität und Vergleichbarkeit werden damit besonders stark gewichtet – Anforderungen, die, wie Wampfler (2024b) argumentiert, nur begrenzt einzulösen sind. Neue Prüfungsformate geraten dadurch häufig in eine asymmetrische Begründungssituation: Während alternative Formate ihre Validität und Bewertbarkeit erst aufwendig legitimieren müssen, gelten etablierte Verfahren als selbstverständlich vergleichbar und rechtssicher.

Auch in pädagogisch reformorientierten Settings zeigen empirische Studien die Wirkmächtigkeit des Selektionsmechanismus. Godau und Kolleg:innen (2025) weisen nach, dass Benotung selbst in kreativitäts- und autonomieorientierten Arrangements „unhintergehbar und omnipräsent“ bleibt und schulische Praxis strukturiert (S. 173). Diskursprägende Leitbilder, die Lehrkräfte primär als Lernermöglicher:innen beschreiben, greifen daher zu kurz. Die strukturelle Wirkmächtigkeit von Benotungspraktiken und die daraus resultierenden Mehrfachrollen des Lehrer:innenhandelns lassen sich nicht normativ ausblenden, sondern müssen als konstitutives Moment schulischer Praxis verstanden werden.

Die Metaanalyse von Breidenstein (2018) zum „Theorem der schulischen Selektionsfunktion“ kritisiert die im Diskurs häufig anzutreffende vorschnelle Auslagerung des Selektionsauftrags an „die Gesellschaft“. Eine solche Perspektive verlagert Selektionsentscheidungen auf eine abstrakte Systemebene und läuft Gefahr, die konkrete pädagogische und didaktische Praxis schulischer Leistungsbewertung zu entlasten. Wird Selektion allein als externer Auftrag verstanden, geraten jene schulischen Routinen, Bewertungsmaßstäbe und Entscheidungsspielräume aus dem Blick, in denen Selektionswirkungen tatsächlich hervorgebracht werden.

Für die Analyse schulischer Prüfungen folgt daraus, Selektionswirkungen nicht lediglich vorauszusetzen, sondern sie im konkreten Vollzug von Unterricht und Prüfung zu rekonstruieren. Auch dort, wo Prüfungen wie das Abitur explizit den Zugang zu Studium und Beruf regulieren, werden Selektionsentscheidungen nicht außerhalb der Schule getroffen, sondern in schulischer Praxis produziert. Aufgabenformate, Bewertungsverfahren und institutionelle Routinen sind daher nicht bloßer Vollzug gesellschaftlicher Anforderungen, sondern ein in der Schule selbst konstituierter Bestandteil der Prüfungskultur. Deren explizite und implizite Wirkmächtigkeit darf bei der Konzeption neuer Prüfungsformate nicht ausgeblendet werden, insbesondere dann nicht, wenn Leistung im Sinne eines *Assessment for Learning* stärker in den Dienst des Lernens gestellt werden soll (Black & Wiliam 1998).

3.2 Grenzen traditioneller Prüfungsformate

Konventionelle Prüfungsformate, insbesondere standardisierte Abschlussprüfungen, stoßen bei der Erfassung komplexer Lernleistungen an deutliche Grenzen. Standardisierte Wissenstests ermöglichen zwar objektive und für Lehrkräfte gut handhabbare Leistungsbewertungen, erfassen aber überwiegend reproduktives Wissen und grundlegende kognitive Fertigkeiten. Anspruchsvollere Fähigkeiten wie das Entwickeln kreativer Lösungswege und neuartiger Problemlösungen, auch in kollaborativen Arbeitsprozessen, lassen sich auf diese Weise kaum abbilden. Darling-Hammond und Adamson (2010) zeigen entsprechend, dass viele traditionelle Tests nicht geeignet sind, höherwertige Fähigkeiten von Schüler:innen angemessen zu erfassen. Vor allem summative Assessments, die ausschließlich das Endprodukt bewerten, berücksichtigen die kognitiven Verfahren und Strategien, die zu diesem Produkt führen, nur unzureichend (Griffin & Care 2015). Ein Übermaß solcher Tests und Prüfungen kann zudem negative Rückwirkungen auf den Unterricht entfalten. Studien zeigen, dass lehrstoffzentrierte Prüfungen Lehrkräfte davon abhalten können, anspruchsvollere kognitive Ziele (*higher-order thinking*) zu verfolgen, weil vor allem abgefragtes Faktenwissen ins Zentrum rückt. *Teaching to the Test* verengt dann den Lernfokus und steht der Förderung komplexer Kompetenzen entgegen (Conley & Darling-Hammond 2013; Foster & Piacentini 2023).

Ein weiteres Problem traditioneller Klausuren und Abschlussprüfungen ist ihre geringe Authentizität. Häufig handelt es sich um isolierte Einzelleistungen unter recht künstlichen Bedingungen: stilles Bearbeiten, Alleinarbeit, Verzicht auf technische und andere Hilfsmittel sowie künstlicher Zeitdruck. Mit realen Anwendungs- und Leistungssituationen im Beruf und im Alltag jenseits der Schule haben solche Prüfungen nur wenig zu tun. Swiecki und Kolleg:innen (2022) kritisieren traditionelle Prüfungen entsprechend als aufwendig, gleichförmig, veraltet und wenig realitätsnah und regen an, KI-Technologien zur Überwindung dieser Defizite zu nutzen. Auch Hurrelmann (2020) argumentiert mit Blick auf das deutsche Abitur, dass der gegenwärtige „Abi-Modus“ den Qualifikationsanforderungen des 21. Jahrhunderts nicht gerecht werde. Junge Menschen benötigten im weiteren Bildungs- und Berufsleben Fähigkeiten, die über reines Faktenwissen hinausgehen; genau diese würden durch das klassische Prüfungssystem jedoch nur unzureichend erfasst (ebd.). Zudem erzeugt die Abbildung komplexer Kompetenzen wie Kreativität oder Kommunikation in standardisierten Schultests und Klausuren ein Validitätsproblem. So würde man etwa keiner Person allein nach bestandener Theorieprüfung bereits die Kompetenz zum Autofahren zuschreiben, sondern erst in Kombination mit einer praktischen Anwendung (Darling-Hammond & Adamson 2010). Die COVID-19-Pandemie mit den langen Schulschließungen und den hohen Belastungssituationen von Kindern und Jugendlichen wirkte wie ein Brennglas: Sie machte Schwächen und Limitationen des bisherigen Prüfungsformats sichtbar und eröffnete zugleich Diskussionen über reformorientierte Ansätze (Hurrelmann 2020; Sliwka & Klopsch 2022).

Schulische Prüfungen stehen unter der Spannung eines grundlegenden Funktionswiderspruchs: Sie sollen diagnostisch Lernprozesse analysieren und fördern und entscheiden zugleich als Selektionsinstrument über Bildungs- und Berufswege.

Zusammenfassend zeigt sich, dass schulische Prüfungen unter der Spannung eines grundlegenden Funktionswiderspruchs stehen: Sie sollen diagnostisch Lernprozesse analysieren und fördern und entscheiden doch zugleich als Selektionsinstrument über Bildungs- und Berufswege. Traditionelle Prüfungsformate besitzen zwar eine historisch gewachsene institutionelle Legitimation und gelten als gut machbar, schnell vergleichbar sowie rechtssicher; sie erfassen jedoch überwiegend reproduktives Wissen und stoßen bei der Messung komplexer Kompetenzen an deutliche Grenzen. Daraus entsteht eine paradoxe Situation: Prüfungen behalten trotz begrenzter Validität und teilweise problematischer Fairness eine hohe normative Stabilität, während gerade jene Kompetenzen, die in aktuellen Bildungsdiskursen als zentral gelten, im bestehenden Prüfungssystem nur unzureichend sichtbar werden.

Dies hat in den letzten 10 bis 15 Jahren in verschiedenen Bildungssystemen weltweit zu verstärkten Reformbestrebungen geführt, neue, zeitgemäßere Formen der Leistungsbewertung zu entwickeln. Die folgenden internationalen Ansätze zeigen, wie Bildungssysteme versuchen, auf diese Herausforderungen zu reagieren.



4 Internationale Ansätze und Forschungsstand zur Leistungsbewertung komplexer Kompetenzen

Die Forschungs- und Schulpraxis hat auf die genannten Spannungsfelder reagiert, indem in vielen Schulsystemen alternative Assessment-Formate entwickelt und erprobt wurden. Im Folgenden werden zentrale Ansätze vorgestellt, die das Ziel haben, komplexe, mehrschichtige Lernergebnisse angemessener und genauer zu erfassen und zu dokumentieren.

4.1 Voraussetzungen für zeitgemäßes Prüfen

Zentrale Voraussetzung für zeitgemäßes Prüfen ist das sogenannte *Constructive Alignment* (Biggs 2014). Gemeint ist die systematische Abstimmung von Lernzielen, Lernaktivitäten und Prüfungsformen. Prüfungen sollen also genau jene Kompetenzen erfassen, die zuvor als Lernziele definiert und im Unterricht gezielt gefördert wurden. Dabei sollten Prüfungen evidenzbasiert gestaltet werden (Brandt et al. 2025; Mislevy et al. 2004). Das bedeutet erstens, dass eine präzise Vorstellung davon entwickelt wird, welche Kompetenz erworben werden soll und wie sie sich entlang eines Kontinuums von Anfänger:innen bis Expert:innen entwickelt. Zweitens ist zu definieren, wie sich diese Kompetenz in beobachtbaren Handlungen von Schüler:innen manifestiert. Drittens werden auf dieser Grundlage Lernsettings und Aufgaben entworfen (*Instructional Design*), die geeignet sind, dieses erwünschte Verhalten hervorzubringen. Empirische Metaanalysen zeigen, dass selbst bei gut erforschten Konstrukten wie kritischem Denken erhebliche Unterschiede in der Operationalisierung bestehen. Gerade deshalb kommt *Constructive Alignment* in Verbindung mit *Evidence-Centered Design* bei der Messung komplexer Kompetenzen eine hohe Bedeutung zu (Abrami et al. 2015).

Das Center for Assessment in New Hampshire rät vor dem Hintergrund des aktuellen Forschungsstandes davon ab, vorschnell Bewertungsraster für die Vergabe von Zensuren anzulegen. Voraussetzung hierfür wäre, dass die zugrunde liegenden Teilkompetenzen empirisch fundiert und domänenspezifisch erforscht sind. Je stärker folgenreiche Tests standardisiert werden, wie es etwa im Abitur oder in anderen *Large-Scale Assessments* der Fall ist, desto höher sind die Anforderungen an die Gütekriterien der Messverfahren (Brandt et al. 2025).

4.2 Kollektive Wirksamkeit von Lehrkräften *(Collective Teacher Efficacy)*

Eine weitere wirksame Rahmenbedingung für zeitgemäßes Prüfen scheint die sogenannte *Collective Teacher Efficacy* zu sein. Gemeint ist die geteilte Überzeugung von Lehrkräften, durch gemeinsames professionelles Handeln wirksam Einfluss auf Lernprozesse und Lernergebnisse nehmen zu können (Donohoo et al. 2020). Empirische Studien ordnen diesem Konstrukt eine besonders hohe Effektstärke auf den Lernerfolg zu und verorten es damit unter den einflussreichsten Wirkfaktoren schulischen Lernens (Hattie 2023). Entscheidend ist jedoch, dass kollektive Wirksamkeitsüberzeugungen nicht isoliert entstehen, sondern an bestimmte organisationale Bedingungen gebunden sind: Anderson und Kolleg:innen (2023) zeigen, dass insbesondere gemeinsame Zielklarheit, die systematische Nutzung von Evidenzen aus Schüler:innendaten sowie unterstützende Führungsstrukturen zentrale Voraussetzungen darstellen. Ontario in Kanada kann hier als exemplarischer Kontext gelten, in dem diese Bedingungen frühzeitig gezielt aufgebaut wurden. So zum Beispiel durch die institutionalisierte Praxis der *Collaborative Teacher Inquiry*, die professionelle Zusammenarbeit, datenbasierte Reflexion sowie Unterrichts- und Prüfungsentwicklung eng miteinander verzahnt (Ontario Ministry of Education 2010). Innovative Prüfungsformate sind somit auf kollektive professionelle Schulkulturen angewiesen; ihre Wirksamkeit hängt wesentlich von Zusammenarbeit und geteiltem Verantwortungsverständnis im Kollegium ab.

4.3 Authentische und performanzorientierte Prüfungen

Einer der zentralen innovativen Ansätze ist die performanzorientierte Leistungsbewertung (*Performance-Based Assessment*). Dabei bearbeiten Schüler:innen authentische Aufgabenstellungen, die offen, komplex und an realen Anwendungssituationen orientiert sind. Beispiele sind umfangreiche Projekte, die etwa Arbeit in komplementären Teams, Forschungsaufgaben, Laborexperimente, Zwischen- und Abschlusspräsentationen oder die Erstellung authentischer Produkte umfassen. Ziel ist, dass Lernende ihr Wissen generativ einsetzen, also etwa eine Forschungsfrage in Biologie experimentell untersuchen, eine historische Debatte im Rollenspiel führen, ein Produkt entwerfen und bauen oder eine komplexe Fallstudie unter Rückgriff auf Wissen aus mehreren Fächern im Team lösen. Solche Aufgaben verlangen nicht nur Fachwissen, sondern ein breites Bündel an unterschiedlichen komplexen Kompetenzen.

Anspruchsvollere Aufgaben dieser Art sind in einigen Schulsystemen bereits fest verankert: Hochleistungsschulsysteme verwenden zunehmend offene, anspruchsvolle Prüfungsaufgaben, die umfangreiches Schreiben, Recherche sowie die Anwendung von Wissen in neuen Situationen erfordern (Darling-Hammond & Adamson 2010). Hier zeigt sich Constructive Alignment in besonderer Klarheit: Mehrdimensionale Unterrichtsziele entsprechen dem, was dann auch in der Prüfung verlangt wird. Wenn Problemlösen und kritisches Denken Lernziele sind, müssen Prüfungen komplexe Problemaufgaben enthalten, die Offenheit in Lösungswegen und -strategien zulassen.

Performance-Based Assessments stellen jedoch erhöhte Anforderungen an Durchführung und Bewertung. Sie müssen häufig durch fachkundige Lehrpersonen beobachtet und anhand strukturierter Bewertungsraster beurteilt werden; im englischsprachigen Diskurs spricht man hier von *Rubrics*. Dieses Vorgehen gilt derzeit noch als arbeitsaufwendig und fehleranfällig. Es eignet sich daher vor allem für intrapersonelle Entwicklungsdiagnostik und lerndiagnostische Verfahren und bleibt hinsichtlich seiner Vergleichbarkeit anfällig für Messfehler.

Zugleich hat PISA (OECD 2017) mit der Operationalisierung komplexer Kompetenzen wie kreativem Denken wichtige Vorarbeiten geleistet; die entsprechenden Verfahren können aus messtheoretischer Sicht jedoch erst einen Anfang darstellen, unter anderem, weil das Konstrukt erstmals erhoben wurde, auf einem vergleichsweise kleinen Itempool basiert und die psychometrische Robustheit sowie kulturelle Vergleichbarkeit der Messung noch begrenzt sind (Rutkowski & Rutkowski 2025).

Vor diesem Hintergrund plädiert die Forschung des Learning Policy Institute für sogenannte *Instructionally Relevant Assessment Systems*, in denen Prüfungen nicht isoliert als reine Messungen verstanden werden, sondern als integraler Bestandteil von Lehr- und Lernprozessen (Badrinarayan 2024). Solche Systeme kombinieren unterschiedliche Aufgabenformate, etwa *Performance-Based Assessments* und Portfolios, und sollen Lehrkräften verschiedene Informationen liefern, die dann auch unmittelbar für Unterrichtsentscheidungen formativ nutzbar sind und die Anpassungsfähigkeit des Unterrichts unterstützen (Badrinarayan 2025). Grundlage sind Designprinzipien wie authentische Aufgaben, eine enge Anbindung an das Curriculum sowie entwicklungsorientierte formative Rückmeldungen, die Lernfortschritte entlang von *Learning Progressions* stufenweise sichtbar machen. Leistungsbewertung soll hier nicht nur summativ am Ende eines Prozesses, sondern bereits formativ während des Lernens wirksam werden.

Assessment for Learning: Lernende erhalten bereits während des Arbeitsprozesses reichhaltiges Feedback und können so Fähigkeiten der Selbstreflexion und der Selbstregulation entwickeln.

Bewährt hat sich zudem ein abgestimmtes Training der Lehrkräfte: Sie korrigieren beispielhafte Arbeiten gemeinsam, kalibrieren ihre Maßstäbe und erhöhen so die Vergleichbarkeit der Bewertungen komplexer Kompetenzen. Auf diese Weise wird die *Assessment Literacy*, also die Bewertungs- und Rückmeldekompetenz von Lehrkräften gestärkt. In einigen Systemen, wie etwa im australischen Bundesstaat Queensland oder in Schulen des New York Performance Standards Consortium, werden Lehrkräftebewertungen zusätzlich durch externe Audits oder Zweitkorrekturen abgesichert, um die Gütekriterien der Messverfahren zu verbessern (Conley & Darling-Hammond 2013). Trotz des Aufwands zeigen Studien, dass gut designte *Performance-Based Assessments* validere Prädiktoren für Studien- und Berufserfolg sein können als traditionelle Prüfungsformen, weil sie ein breiteres Fähigkeitsspektrum abbilden (ebd.) und der Komplexität heutiger beruflicher Anforderungen eher entsprechen. Zugleich fokussieren sie auf *Assessment for Learning*: Lernende erhalten bereits während des Arbeitsprozesses reichhaltiges Feedback und können so Fähigkeiten der Selbstreflexion und der Selbstregulation entwickeln. Genau diese Kompetenzen sollen sie nach der Schule dazu befähigen, ihr Lernen eigenständig weiter zu reflektieren und zu steuern.

4.4 Formatives Assessment und Lernprogressionen komplexer Kompetenzen

Neben summativen Abschlussprüfungen rückt also das formative Assessment mittlerweile deutlich stärker in den Fokus, wenn komplexe Kompetenzen aufgebaut und dokumentiert werden sollen. Formative Assessments sind laufende, von Beginn an in den Lernprozess integrierte Beurteilungsformen, die nicht primär der abschließenden Bewertung, sondern vor allem der Lernförderung dienen durch Feedback von Expert:innen, *Peer Assessment* oder angeleitete und strukturierte Selbstevaluation und -reflexion.

Zunehmend ist davon auszugehen, dass auch KI-basierte Tools eingesetzt werden, um formatives Feedback unmittelbar und personalisiert zu ermöglichen (siehe → Kapitel 6). Solche Systeme adressieren ein strukturelles Grundproblem traditioneller Schulbildung: Eine einzelne Lehrperson kann angesichts großer Lerngruppen die Anforderungen an wirksames Feedback oft nur begrenzt erfüllen. Wirksames Feedback muss jedoch zeitnah erfolgen und sich diagnostisch am Lernstand sowie an der sog. Zone der nächsten Entwicklung einzelner Schüler:innen oder Gruppen orientieren (Vygotsky 2021; Helmke 2022).

In vielen Ländern werden mittlerweile digitale Portfolios eingesetzt, in denen Schüler:innen einzeln oder im Team über längere Zeiträume Arbeiten sammeln, die ihre Kompetenzentwicklung dokumentieren, von Projektberichten bis zu Präsentationsvideos (López-Crespo et al. 2022). Solche

Portfolios können in Abschlussjahrgängen selbst Prüfungsleistungen darstellen, oft kombiniert mit einem mündlichen Kolloquium oder einer Präsentationsprüfung, in der die Lernenden ihr Portfolio erläutern und reflektieren.

Ansätze im Sinne eines *Assessment for Learning* tragen der Tatsache Rechnung, dass Lernende Kompetenzen nicht im Gleichschritt erwerben. Deshalb gibt es zunehmende Bestrebungen, Meilensteine und Lernfortschritte personalisiert und differenziert zu dokumentieren und zu bewerten. In der Forschung wird hier vom Konzept der *Learning Progressions* gesprochen. Zwar liegen bislang noch keine durchgängig wissenschaftlich fundierten Lernverlaufsmodelle vor (Brandt et al. 2025), doch arbeiten Bildungssysteme international bereits häufig an der Definition von Kompetenzstufen für überfachliche und komplexe Kompetenzen, so etwa Ontario / Kanada für *Global Competencies* (Ontario Ministry of Education 2024b), die Niederlande für *breden vaardigheden* (überfachliche Kompetenzen; Curriculum.nu 2018) oder Australien im *F-10 Curriculum* (Australian Curriculum, Assessment and Reporting Authority 2022). In der Praxis nutzen Schulen häufig von Lehrkräften entwickelte Kompetenzraster, um sowohl den aktuellen Entwicklungsstand als auch weiteres Entwicklungspotenzial abzubilden. Entscheidend ist, dass solche Beurteilungen transparente Kriterien und eine gemeinsame Sprache für Lehrende und Lernende bereitstellen. Nur so verstehen Schüler:innen, woran sie arbeiten, wie sie über ihr eigenes Lernen nachdenken und sprechen können und wie sie sich in Bezug auf eine bestimmte Kompetenz weiterentwickeln können.

4.5 Leistungsbewertung in kooperativen Arbeitsformen

Eine besondere Herausforderung stellt die Leistungsbewertung in kollaborativen Settings dar. In Lernformen wie *Deeper Learning*, *Project-Based Learning* oder *Problem-Based Learning* arbeiten Schüler:innen häufig in Teams an gemeinsamen Aufgaben (Sliwka & Klopsch 2022). Dadurch trainieren sie Kompetenzen des kollaborativen Problemlösens, die explizit gefördert werden sollen. Damit stellt sich unmittelbar die Frage, wie der Beitrag Einzelner zu einer Gruppenleistung erfasst werden kann. Dieser anspruchsvollen Problematik widmet sich inzwischen eine wachsende Zahl von Forschungsarbeiten. Kollaboratives Problemlösen (*Collaborative Problem Solving*, CPS) gilt mittlerweile als eigenständige Kompetenz, die kognitive und soziale Fähigkeiten miteinander verbindet (Hesse et al. 2015; OECD 2017).

International für Aufsehen sorgte 2015 die Einführung eines CPS-Tests in PISA. Es handelte sich um den ersten politisch sichtbaren und methodisch ambitionierten internationalen Versuch, die Fähigkeit 15-Jähriger zum Problemlösen im Team zu messen. Aufgrund praktischer Beschränkungen interagierten die Schüler:innen jedoch nicht mit realen Partner:innen,

sondern mit Computersimulationen: In einem computerbasierten Test traten simulierte Teammitglieder als virtuelle Agenten auf, die über Chat mit den Teilnehmenden interagierten. Die Schüler:innen mussten mit diesen Agenten kommunizieren, Informationen austauschen, Konflikte bearbeiten und gemeinsam Lösungen entwickeln. Die Aufgaben umfassten unter anderem Gruppenentscheidungen, kooperative Informationsweitergabe und gemeinsame Produktentwicklung. Das Verfahren zeigte, dass die Messung kollaborativer Fähigkeiten grundsätzlich möglich ist, allerdings auch, dass sie indirekt bleiben muss, sobald Individualdaten benötigt werden, die den Gütekriterien der Leistungsmessung entsprechen.

Die Reflexion von Graesser und Kolleg:innen (2018) macht deutlich, dass diese Form der Messung mit grundlegenden Spannungen verbunden ist: Kollaboration wird auf individuelle Performanz reduziert, obwohl sie theoretisch als gruppenbezogenes und komplexes emergentes Phänomen verstanden wird, also in gewisser Weise *messy* ist. Die Nutzung computerbasierter Agenten ermöglicht zwar Standardisierung und Vergleichbarkeit, strukturiert die Interaktion aber zugleich stärker vor als in einer echten Kollaborationssituation. Kollaboration wird dadurch primär über beobachtbare Kommunikationshandlungen erfasst. Vielschichtige und nicht vollständig antizipierbare soziale Interaktionen bleiben nur begrenzt sichtbar, weil Aufgaben Kollaboration gezielt inszenieren und vorhersehbar machen müssen, um sie in eine messbare Form zu überführen.

PISA 2015 zeigte in diesem Testbereich deutliche Leistungsunterschiede zwischen Ländern und lieferte Hinweise darauf, welche Bildungssysteme Teamfähigkeit besonders fördern (OECD 2017). Neuere Forschungsprojekte versuchen daher, echte Gruppenarbeiten genauer zu analysieren, etwa mithilfe von *Learning Analytics*: Digitale Chatverläufe, Kommunikationsmuster oder gemeinsame Artefakte von Schüler:innengruppen werden mit KI-Methoden ausgewertet, um Indikatoren erfolgreicher Zusammenarbeit zu identifizieren (Koh et al. 2023; Ouyang et al. 2023). Gerade Studien zu KI-gestützten Systemen zeigen, wie Kollaboration zunehmend über wiederkehrende multimodale Interaktionsmuster erfasst werden kann – also nicht nur über einzelne Beiträge einzelner Lernender, sondern über deren Abfolgen und zeitliche Dynamiken, etwa den Wechsel zwischen Planung, Koordination und Ausführung. Die Skalierung solcher Ansätze und die methodische Verbindung individueller Beiträge mit emergenten Gruppenprozessen bleiben jedoch zentrale Herausforderungen innovativer Formate der Messung komplexer Kompetenzen (Ouyang et al. 2023).

In der Schulpraxis werden zur Bewertung von Gruppenarbeiten gegenwärtig verschiedene pragmatische Lösungen eingesetzt, die Schüler:innen und Lehrkräfte allerdings oft noch als unbefriedigend erleben und die teilweise widersprüchliche Handlungsanforderungen erzeugen (Helmke 2022). Häufig wird die Gruppenleistung als Ganzes bewertet, etwa die Qualität eines Produkts oder einer Präsentation, während individuelle Beiträge zusätzlich über *Peer Assessment* und Selbstevaluation sowie mündliche Prüfungen erfasst werden. So können etwa alle Schüler:innen ein Lerntagebuch führen oder eine Reflexion über den eigenen Anteil verfassen, die in die Note einfließt. Auch Peer Assessment kann Hinweise darauf geben, wer in welcher Weise zum Gruppenerfolg beigetragen hat (Topping 2009).

Entscheidend ist dabei, Teamarbeit selbst als Lernziel zu verstehen und systematisch in die Bewertung einzubeziehen. Einige Schulen beurteilen daher auch die Kompetenz zur Zusammenarbeit anhand von Kriterien wie Kommunikationsfähigkeit, Rollenübernahme und Konfliktlösung im Team. Dies setzt klare Kriterien voraus; andernfalls bleibt die Bewertung von Zusammenarbeit subjektiv und potenziell ungerecht. Die Forschung legt nahe, für wirksame Zusammenarbeit relevante Teilkompetenzen möglichst beobachtbar zu machen, etwa durch Zwischenpräsentationen, Einzel- und Team-Interviews oder die Analyse von Interaktionsmustern, und individuelle Lernerträge im Zweifel zusätzlich separat zu prüfen, beispielsweise durch einen kurzen schriftlichen Test nach Abschluss der Gruppenarbeit. Diese Vorgehensweise entspricht aktuellen internationalen Empfehlungen, die Kollaboration als Bündel beobachtbarer Teilkompetenzen verstehen und zur Sicherung von Validität und Fairness eine Kombination aus Gruppen- und Individualleistungen empfehlen (MBL Work Group 2021).

4.6 Bewertung kreativen Denkens und weiterer komplexer Kompetenzen

Neben Kollaboration gelten insbesondere Kreativität und kritisches Denken als schwer messbare, aber zentrale Kompetenzen. Auch hier gibt es inzwischen Initiativen, geeignete Prüfungsformate zu entwickeln. Ein prominentes Beispiel ist PISA 2022, in dem kreatives Denken erstmals als internationaler Testschwerpunkt enthalten war. Schüler:innen bearbeiteten offene Aufgaben, etwa das Weiterentwickeln einer abstrakten Zeichnung oder das schriftliche Formulieren kreativer Ideen zu einem Szenario; bewertet wurde anschließend anhand definierter Kreativitätskriterien (Foster & Piacentini 2023). Erste Auswertungen deuten darauf hin, dass solche Tests verschiedene Dimensionen kreativen Denkens, wie zum Beispiel Ideenreichtum und Ideenausarbeitung, erfassen können, zugleich aber stark kulturell beeinflusst sind. Forschungsergebnisse legen nahe, dass Jugendliche aus unterschiedlichen kulturellen Kontexten Kreativität zum Teil unterschiedlich definieren. In einigen ostasiatischen Kulturen wird etwa Nützlichkeit stärker betont als Originalität (Brandt et al. 2025).

Als wirksames Verfahren wird die Methode des *Adaptive Comparative Judgement* als geeignet empfunden. Anstelle starrer Kompetenzraster werden dabei Schüler:innenarbeiten oder Teamprodukte paarweise verglichen: Lehrkräfte entscheiden wiederholt, welche von zwei Arbeiten beispielsweise kreativer ist oder argumentativ stärker ausfällt. Aus einer Vielzahl solcher Vergleiche entsteht schrittweise eine Rangordnung, aus der mithilfe statistischer Modelle konsistente Bewertungen abgeleitet werden können. Pilotstudien bescheinigen diesem Verfahren bei der Beurteilung kreativer Leistungen eine vielversprechende Reliabilität (Pollitt 2012; Bramley 2015). In einigen britischen Schulen wird *Adaptive Comparative Judgement* bereits

genutzt, um projektartige Prüfungen zu bewerten, weil sich damit schwer objektivierbare Qualitätsunterschiede komplexer Arbeiten vergleichsweise effizient erfassen lassen.

Zur Konkretisierung des Diskurses lässt sich zudem der *Cornell Critical Thinking Test* nennen, der als standardisiertes Instrument zur Erfassung kritischen Denkens für die 11. und 12. Klasse entwickelt wurde (Leach et al. 2020). Der Test umfasst überwiegend Multiple-Choice-Aufgaben, in denen Schüler:innen etwa Schlussfolgerungen bewerten, Annahmen identifizieren oder Argumente analysieren müssen. Ziel ist es, zentrale Teilfähigkeiten kritischen Denkens wie Induktion, Deduktion und Evidenzbewertung möglichst objektiv und zuverlässig zu erfassen. Gerade an diesem Beispiel zeigt sich jedoch, dass die Operationalisierung kritisches Denken auf kognitiv-diskursive Teilaspekte reduziert, während kontextgebundene, kreative oder sozial eingebettete Dimensionen nur begrenzt berücksichtigt werden (Barta et al. 2024).

Bildungssysteme und Wissenschaft bewegen sich in einem Feld mit vielen offenen Fragen zu den Gütekriterien neuer Prüfungsformate.

Eine aktuelle *Scoping-Review*-Studie von Kain und Kolleg:innen (2024) zeigt, dass zur Messung dieser Kompetenzen bislang vergleichsweise wenig Forschung vorliegt und viele Arbeiten eher die Umsetzung von Unterrichtsformaten als die Prüfungsformate untersuchen, was möglicherweise darauf schließen lässt, dass erst Lernsettings und dann Prüfungsformate entwickelt werden. Dies unterstreicht den experimentellen Charakter zahlreicher aktueller Reformen: Bildungssysteme und Wissenschaft bewegen sich hier in einem Feld mit vielen offenen Fragen zu den Gütekriterien neuer Prüfungsformate. Vor allem bei Prüfungen im Rahmen wichtiger Bildungsübergänge und Abschlüsse bremst die Sorge um juristische Legitimität, Ressourcenökonomie und gesellschaftliche Akzeptanz das ambitionierte Policy-Design derzeit noch aus.



5 Internationale Praxisbeispiele zur Leistungsbewertung in Abschlussprüfungen

Weltweit gibt es dennoch bereits unterschiedliche Ansätze, wie solche komplexen Schüler:innenleistungen insbesondere in Abschlussprüfungen bewertet und geprüft werden. Im Folgenden werden einige Länder und Initiativen vorgestellt, die als praktische Beispiele dienen und die Diskussion in Deutschland bereichern können.

5.1 Finnland, Dänemark und Norwegen

Die nordischen Länder werden häufig als Beispiele progressiver Bildungsreformen genannt; das gilt auch für den Bereich des Assessments. Finnland hat sein modulares Curriculum in den vergangenen Jahren umfassend kompetenzorientiert weiterentwickelt. Seit der Curriculumreform von 2016 gibt es sieben transversale Kompetenzen (*laaja-alainen osaaminen*), die alle Fächer durchziehen, darunter Denken und Lernen lernen, kulturelle und kommunikative Kompetenz, Selbstkompetenz, digitale Kompetenz und Citizenship. Selektionswirksame standardisierte Tests werden dort im Wesentlichen auf die Abschlussprüfung am Ende der Sekundarstufe II begrenzt, während im Unterricht lehrkraft- und peer-basierte formative Assessment-Ansätze stark betont werden (Eurydice Finland 2025). Das reduziert den Prüfungsdruck und stärkt kontinuierliches Lernen und Selbstregulation gegenüber externer Leistungsbeurteilung.

Zugleich werden vermehrt interdisziplinäre Projekte im Sinne des *Phenomenon-Based Learning* durchgeführt, in denen Schüler:innen Themen fächerübergreifend und meist projektorientiert bearbeiten. Die Bewertung dieser Kompetenzen erfolgt überwiegend schulisch intern und qualitativ: Lehrkräfte beobachten Lernende in Projekten, geben Rückmeldungen und dokumentieren Kompetenzausprägungen teils narrativ in Zeugnissen. Wichtig ist dabei, dass *Phenomenon-Based Learning* in Finnland nicht als Ersatz der Fächer gedacht ist, sondern als Ergänzung, die fachliche Tiefe mit überfachlichen Perspektiven verbindet (Schaffar & Wolff 2024). Zugleich verweist auch die finnische Debatte auf die Spannung, die entstehen kann, wenn ganzheitliche Bildungsansprüche durch internationale Mess- und Outcome-Logiken unter Druck geraten (ebd.).

In der Sekundarstufe II (gymnasiale Oberstufe) wurde das traditionelle Abitur (*Ylioppilastutkinto*) zwar nicht abgeschafft, aber modernisiert. Dort hat die kriteriale Bezugsnorm (Vergleich mit Kriterien) die ehemals dominante soziale Bezugsnorm (Vergleich mit der Peer-Gruppe) bereits abgelöst und Prüfungsformate dadurch flexibilisiert. Finnlands Abiturprüfungen sind seit 2019 vollständig digital, was zumindest neue Aufgabenformate ermöglicht (z. B. interaktive Aufgaben, Audio- und Videomaterial sowie Datenbanken zur Recherche in Prüfungen). Allerdings bleibt das Abitur eine fachbezogene Prüfung; komplexe überfachliche Fähigkeiten werden dort indirekt mit erfasst, etwa durch Essayfragen, die kritisches Denken verlangen, oder durch fächerübergreifende Konstellationen (so gibt es Module, in

denen z. B. Biologie und Philosophie gemeinsam geprüft werden können, um vernetztes Denken zu fördern). Zudem ist die Oberstufe modular organisiert und nicht an starre Jahrgangsstrukturen gebunden. Der Bildungsgang ist in der Regel auf etwa drei Jahre angelegt, kann jedoch flexibel verkürzt oder verlängert und nach offiziellen Angaben in zwei bis vier Jahren abgeschlossen werden, was den Schulabschluss zeitlich deutlich flexibler organisiert als in stärker getakteten Oberstufensystemen (Finnish National Agency for Education 2026b, 2026c).

Eine interessante Entwicklung in Finnland ist die Messung der Lernkompetenz: In einigen Schulen werden Tests zum *Learning to Learn* eingesetzt (ein Konstrukt, das Problemlösefähigkeit, Metakognition und Motivation umfasst), um Schüler:innen und Lehrkräften Rückmeldung über solche übergreifenden Fähigkeiten zu geben (Hautamäki & Kupiainen 2014). Auch wenn diese Tests nicht im Abitur verankert sind, fließen sie formativer in die Förderplanung ein. Finnland setzt dabei auf nationale Stichprobentests (*sample-based evaluation*) anstatt auf Schulrankings, welche dann unter anderem von FINEEC, dem Finnish Education Evaluation Centre (2026), evaluiert werden.

Darüber hinaus hat das finnische Bildungsministerium im Februar 2026 eine neue Vision für das Bildungssystem veröffentlicht: *Basic Education 2045: For Life – A Vision for Finnish Comprehensive Schools* (Finland Ministry of Education and Culture 2026). Darin wird auch die veränderte Rolle von Assessment thematisiert: Leistungsdiagnostik und Bewertung erscheinen hier vor allem als Mittel zur Förderung des Lernens, zur Stärkung von Chancengerechtigkeit und zur Qualitätsentwicklung des Systems. Im Mittelpunkt steht eine lernunterstützende, faire, transparente und verlässliche Bewertung, die Schüler:innen hilft, ihre Lernprozesse besser zu verstehen und ihre Fähigkeit zum selbstständigen Lernen und zur Selbstregulation auszubauen. Bewertung soll sich daher stärker auf den Lernprozess sowie auf tiefes Verständnis und Denkfähigkeiten konzentrieren, statt nur Ergebnisse des Lernprozesses in den Blick zu nehmen.

Auf Schulsystemebene wird Bewertung mithilfe von digitalen Diagnostik-Tools weiterentwickelt, um Lernschwierigkeiten frühzeitig zu erkennen, gezielte Förderung zu ermöglichen und Chancengleichheit zu verbessern. Gleichzeitig dient diese Art der Bewertung dazu, die Qualität des Bildungssystems insgesamt zu sichern und weiterzuentwickeln. Dabei werden zunehmend auch schwer messbare Kompetenzen wie kreatives Problemlösen und ethische Fähigkeiten stärker berücksichtigt. Im finnischen Diskurs ist wichtig, dass Bewertung die intrinsische Motivation der Lernenden nicht beeinträchtigt oder Lernen auf reine Messbarkeit verengt. Insgesamt wird versucht, die ganze Prüfungskultur sukzessive stärker auf *Assessment for Learning* auszurichten: Lernende sollen durch Assessment unterstützt werden, ihren Fortschritt zu reflektieren, ihre Ziele zu klären und ihr Lernen entsprechend selbst zu regulieren.

Auch andere nordische Länder wie Dänemark und Norwegen haben ihre Abschlussprüfungen in den vergangenen Jahren punktuell reformiert. In Dänemark war in vielen schriftlichen Abiturprüfungen über längere Zeit hinweg der Zugang zum Internet erlaubt, wodurch Aufgaben stärker auf

Informationsbewertung, Quellenkritik und Syntheseleistungen ausgerichtet werden konnten. Mit der breiten Verfügbarkeit generativer KI wurde jedoch im Januar 2026 eine neue Prüfungsverordnung verabschiedet, die den Internetzugang wieder stärker einschränkt: Während schriftlicher Prüfungen durften zwischenzeitlich grundsätzlich nur lokal gespeicherte Materialien verwendet werden. Nun wurde im August 2026 ein Sofortpaket gegen KI-Betrug auf den Weg gebracht, das drei kurzfristige Initiativen umfasst: Erstens müssen alle zu Hause verfassten zu benotenden Leistungen künftig mündlich verteidigt werden. Zweitens werden alle weiterführenden Schulen verpflichtet, Kontrollinstrumente wie Überwachungssoftware sowohl in schriftlichen Prüfungen als auch im Unterricht zu nutzen. Drittens werden die Gymnasien ermutigt, schriftliche Aufgaben verstärkt an der Schule unter kontrollierten Bedingungen bearbeiten zu lassen, um den Arbeitsprozess besser nachvollziehen und gerechter bewerten zu können (Undervisningsministeriet 2026). Hier zeigt sich eine Tendenz Richtung mehr Überwachung zur Sicherstellung der Eigenständigkeit der Arbeiten.

Norwegen verfolgt *Assessment for Learning* im gesamten Schulsystem und betont die Bedeutung formativer Rückmeldung als zentralen Teil von Bewertung. Abschlussprüfungen bestehen dort neben schriftlichen Tests auch aus mündlichen Prüfungen, in denen Schüler:innen Projekte präsentieren oder fächerverbindende Fragen erörtern müssen; Kommunikationsfähigkeit und Reflexionsvermögen fließen dabei in die Note ein. Diese Prüfungen werden jedoch schulisch organisiert und variieren entsprechend, was die Vergleichbarkeit einschränkt.

Statt immer neue Tests einzuführen, wird stärker in die Fortbildungen von Lehrkräften investiert, um konsistente und kompetenzorientierte Beurteilungen durch Lehrkräfte zu sichern.

Insgesamt zeigt sich in den nordischen Ländern ein klarer Trend: Lehrkräfteprofessionalisierung und *Assessment Literacy* gelten als Schlüssel für gutes Assessment. Statt immer neue Tests einzuführen, wird stärker in die Fortbildungen von Lehrkräften investiert, um konsistente und kompetenzorientierte Beurteilungen durch Lehrkräfte zu sichern. In vielen Fächern zählen Projektarbeiten oder Facharbeiten zur Jahresleistung und gehen in die Abschlussbewertung ein. Beispiele sind das umfangreiche *Gymnasiearbete* in Schweden, das einer Seminararbeit an der Uni ähnelt, oder praktische Prüfungen in naturwissenschaftlichen Fächern in Norwegen.

5.2 Estland

Estland hat sich als aufstrebendes Bildungssystem stark an Finnland orientiert und ebenfalls kompetenzorientierte Curricula eingeführt. Die staatlichen Abschlussprüfungen bestehen aus zentralen Tests in Kernfächern, die seit einigen Jahren computerbasiert durchgeführt werden. Dadurch konnten neue Aufgabenformate Einzug halten, etwa interaktive problemorientierte Fragestellungen in den Naturwissenschaften oder Schreibaufgaben mit Nutzung digitaler Hilfsmittel. Zusätzlich misst Estland der digitalen Kompetenz besonderes Gewicht bei: Schüler:innen müssen in gesonderten Tests oder Projekten ihre Informationstechnologie- und Medienkompetenz nachweisen. Interessant ist vor allem Estlands Nutzung von Bildungsdaten: Durch das digitale Prüfungssystem (*Examination Information System*) lassen sich große Datenmengen zu Lösungswegen auswerten, die wiederum zur Verbesserung von Aufgaben und Feedback genutzt werden können (Whyte & Nael 2025). So wird versucht, komplexen Kompetenzen schrittweise gerechter zu werden. In Pilotversuchen werden aktuell zudem kollaborative digitale Prüfungen erprobt, bei denen zwei oder mehr Schüler:innen gemeinsam online eine Aufgabe lösen und ihre Kommunikation aufgezeichnet sowie anschließend beurteilt wird. Bislang laufen diese Erprobungen neuer Prüfungsformate allerdings noch außerhalb des Regelsystems.

5.3 USA

In den Vereinigten Staaten gibt es aufgrund des föderalen Systems keine einheitliche Abschlussprüfung, wohl aber zahlreiche Reforminitiativen auf Bundesstaaten- und Distriktebene, die sich mit der Bewertung komplexer Kompetenzen befassen. Ein prominentes Beispiel sind die *Performance-Based Assessment*-Initiativen einzelner Bundesstaaten. New Hampshire startete mit dem Programm *Performance Assessment of Competency Education* (PACE) einen vom US-Bildungsministerium genehmigten Pilotversuch, in dem traditionelle standardisierte Tests in einigen Jahrgangsstufen durch lokal entwickelte *Performance-Based Assessments* ersetzt wurden. Lehrkräfte entwickeln dort anspruchsvolle, in den Unterricht eingebettete Aufgaben, wie zum Beispiel eigene wissenschaftliche Forschungsprojekte für Schüler:innen, und bewerten sie anhand gemeinsamer *Rubrics*. Erste Ergebnisse zeigen, dass PACE-Schüler:innen in externen Vergleichen wie dem standardisierten Scholastic Assessment Test (SAT) nicht schlechter abschneiden als traditionell geprüfte Schüler:innen, im Unterricht jedoch intensiver an tieferem Verständnis arbeiten (Evans 2019).

Auf nationaler Ebene wurde um 2010 mit der Einführung der *Common Core State Standards* ebenfalls versucht, Prüfungen weiterzuentwickeln. Zwei große Testkonsortien – PARCC und Smarter Balanced – entwickelten neue *Common Core Assessments*, die teilweise performanzorientierte Aufgaben

enthielten. So mussten Schüler:innen in Englisch etwa aus mehreren Quellen ein argumentatives Essay verfassen oder in Mathematik mehrschrittige Anwendungsprobleme lösen. Diese Ansätze wurden jedoch nur teilweise umgesetzt, weil sich einzelne Bundesstaaten zurückzogen. Gleichwohl setzten sie einen wichtigen Impuls: Die Diskussion um sogenannte *Next Generation Assessments* ist seither dauerhaft präsent. Organisationen wie das Learning Policy Institute oder das Aurora Institute propagieren *Balanced Assessment Systems*, in denen klassische Tests, schulische *Performance-Based Assessments* und fortlaufende formative Beurteilungen kombiniert werden. Es zeigt sich im Diskurs, dass für Kompetenzen wie Teamarbeit, kritisches Denken oder Lernkompetenz ein einzelner Test nicht ausreicht; erforderlich seien multiple Prüfverfahren (Conley & Darling-Hammond 2013).

Eine weitere Entwicklung in den USA ist der Einzug von Capstone-Projekten in viele High Schools (College Board 2025). Immer mehr Schulbezirke (z. B. in Rhode Island, Oregon oder New Hampshire) verlangen von Abschlussschüler:innen ein größeres Projekt, das am Ende präsentiert werden muss, sei es eine wissenschaftliche Mini-Studie, ein soziales Engagementprojekt oder eine eigene kleine technische Erfindung. Bewertet werden diese *Capstones* meist von Panels aus Lehrkräften und teilweise externen Partner:innen; sie gehen als eigener Kurs in das *Transcript* (Zeugnis) ein. Damit sollen Merkmale komplexer Kompetenzen wie Selbstständigkeit, Projektmanagement, Präsentationsfähigkeit und reflektiertes Arbeiten nachgewiesen werden. Nationale Aufmerksamkeit erhielt auch die *Advanced Placement (AP) Capstone Initiative* des College Board: Schüler:innen können ein zusätzliches Diplom erwerben, wenn sie mit *AP Seminar* und *AP Research* zwei Kurse absolvieren, in denen forschendes Lernen und eine eigene Forschungsarbeit mit Kolloquium im Zentrum stehen. Selbst in den USA, wo standardisierte SAT-Tests in Mathematik und Englisch eine wichtige Rolle bei der Vergabe von Studienplätzen spielen, wird so Raum für komplexere Leistungen geschaffen.

5.4 Kanada

Auch in Kanada haben mehrere Provinzen begonnen, ihre Prüfungs- und Bewertungspraxis zu modernisieren, um komplexe Kompetenzen besser abzubilden. British Columbia führte 2016 ein neues Curriculum mit sogenannten *Core Competencies* (*Thinking, Communication, Personal & Social*) ein. Parallel dazu wurden die früher umfangreichen Abschlussprüfungen reduziert: High-School-Schüler:innen schreiben nun nur noch standardisierte Prüfungen in grundlegenden sprachlichen und mathematischen Kompetenzen, während fachliche Leistungen stärker im Unterricht bewertet werden. Zudem wird von den Schüler:innen erwartet, zum Abschluss eine Selbstevaluation ihrer *Core Competencies* vorzulegen, die dem Zeugnis beigefügt wird. Das stärkt Reflexionsfähigkeit und lenkt Aufmerksamkeit auf überfachliche Kompetenzen, auch wenn dafür bislang keine eigene offizielle Note vergeben wird.

Ontario hat mit der *Aligning and Building Global Competencies*-Initiative in einzelnen Schulbezirken Pilotprojekte angestoßen, in denen *Global Competencies* (kritisches Denken, Innovation, Lernen lernen, Kollaboration, Kommunikation und *Citizenship*) über Projekte und Portfolios erfasst werden. Die politisch einflussreiche Organisation People for Education hat darüber hinaus mit dem Programm *Measuring What Matters* erprobt, wie sich Fähigkeiten wie Kreativität oder *Citizenship* mittels Beobachtungsbögen und Schüler:innen-Surveys sichtbar machen lassen (Hart 2019). Zwar ersetzen diese Verfahren bislang keine Abschlussprüfungen, sie liefern jedoch detaillierte Kompetenzprofile.

In einigen Provinzen sind praktische Prüfungen Teil der Abschlussbewertung, vor allem in beruflichen Bildungsgängen. So müssen Auszubildende oder Fachoberschüler:innen praktische Aufgaben lösen (z. B. ein Businessprojekt planen oder ein Maschinenbauteil entwerfen), die von Prüfer:innen beobachtet und bewertet werden. Auch die allgemeinbildende Oberstufe folgt dieser Entwicklung, indem etwa in Alberta und Manitoba Schüler:innenunternehmen oder Sozialprojekte im Rahmen von *Entrepreneurship*- oder *Leadership*-Kursen durchgeführt und bewertet werden. Quebec hat ein etwas anderes Schulsystem, aber auch dort gibt es Bewegung: Sekundarstufe-II-Schüler:innen präsentieren sogenannte *Personal Orientation Projects*, in denen sie eigene Lernziele verfolgt haben.

Auffällig in Kanada ist der starke Fokus auf Selbstreflexion und Schüler:innenbeteiligung: Lernende werden häufig aktiv in Bewertungsprozesse einbezogen, etwa über Selbsteinschätzungen oder Konferenzen mit Lehrkräften, in denen sie über ihre eigene Entwicklung sprechen. Diese Formate (z. B. *Voice & Choice*; Learner Engagement) sind nicht nur begleitende Angebote, sondern werden als praktischer Teil der Kompetenzentwicklung im Bereich der Selbstregulation gesehen. Zugleich wird die Verzahnung von formativem und summativem Assessment betont: Kompetenzen sollen zunächst im Unterricht eingeführt und sukzessive mit formativem Feedback eingeübt werden, bevor sie in anspruchsvollen Aufgaben summativ nachgewiesen werden können. Kanada zeigt zudem, wie wichtig gesellschaftliche Akzeptanz ist. Der Diskurs um komplexe Kompetenzen beziehungsweise *21st Century Skills* ist dort breit verankert, sodass auch Universitäten und Arbeitgeber:innen solche neuen Bewertungsformen eher anerkennen. Einige kanadische Universitäten akzeptieren mittlerweile portfolio-basierte Kompetenznachweise als Teil der Bewerbungsunterlagen.

Kanada zeigt zudem, dass schulische Abschlusslogiken flexibler organisiert werden können, wobei es allerdings Unterschiede zwischen den Provinzen gibt. In Alberta ist der *High School Completion* ausdrücklich als System mit mehreren Abschlusswegen angelegt. Die Provinz verweist auf verschiedene Abschlussoptionen, darunter das *Alberta High School Diploma*, das *Certificate of School Completion* sowie das *High School Equivalency Diploma*. Zugleich betont Alberta flexible Lernumgebungen, in denen Lernen in unterschiedlichen Settings stattfinden kann und Schüler:innen stärker mitbestimmen, wo, wann und in welcher Form sie lernen (Alberta Ministry of Education 2026b, 2026a).

Für das *Ontario Secondary School Diploma* müssen für den Abschluss 40 Stunden Sozialdienst und zwei Online-Lernkurse absolviert werden.

Ontario kombiniert demgegenüber klar definierte Abschluss-Anforderungen mit zusätzlichen, über den Fachunterricht hinausgehenden Bedingungen. Für das *Ontario Secondary School Diploma* müssen beispielsweise für den Abschluss mindestens 40 Stunden Sozialdienst absolviert werden sowie für Schüler:innen, die nach 2020/21 in die 9. Klassenstufe eingetreten sind, mindestens zwei Online-Lernkurse neben dem Präsenzunterricht belegt werden (Ontario Ministry of Education 2020, 2024a, 2022a). Bemerkenswert ist dabei, dass der Nachweis der sozial geleisteten Arbeit im Zeugnis dokumentiert wird. Damit wird der Schulabschluss nicht allein an fachliche Leistungen gebunden, sondern auch an digitale Lernanteile und an Formen zivilgesellschaftlicher Beteiligung (Ontario Ministry of Education 2022b).

5.5 Australien und Neuseeland

In Australien spielt die Australian Curriculum, Assessment and Reporting Authority (ACARA) eine zentrale Rolle bei der Entwicklung von Curriculum- und Assessmentstrukturen. Neben standardisierten Vergleichstests wie NAPLAN verfolgt Australien seit der Curriculumreform von 2010 einen erweiterten Kompetenzansatz: Das nationale Curriculum enthält sogenannte *General Capabilities*, darunter *Critical and Creative Thinking*, *Personal and Social Capability*, *Ethical Understanding* sowie *Intercultural Understanding*. Diese Kompetenzen sollen fächerübergreifend entwickelt und bewertet werden. Für mehrere dieser überfachlichen Bereiche wurden zudem *Learning Progressions* entwickelt, die beschreiben, wie sich Fähigkeiten über verschiedene Entwicklungsstufen domänenspezifisch entfalten und diagnostizieren lassen. Die Bewertung erfolgt überwiegend formativ und schulisch intern, etwa über projektbasierte Aufgaben, Portfolios oder qualitative Beobachtungen. Ergänzend werden Selbst- und *Peer Assessment* genutzt, um metakognitive Fähigkeiten zu stärken (Australian Curriculum, Assessment and Reporting Authority 2026).

Eine wichtige Rolle in der Weiterentwicklung von Assessment-Modellen spielt die Initiative *Melbourne Metrics* an der University of Melbourne (University of Melbourne, o. J.). Im Rahmen des Programms *New Metrics* werden neue Verfahren entwickelt, um komplexe Kompetenzen wie Kreativität, Problemlösefähigkeit oder Kollaboration zu erfassen, etwa durch projektbasierte Aufgabenformate oder Kompetenz-Profile (Care et al. 2019; University of Melbourne 2025).

Auch Neuseeland verfolgt seit einigen Jahren einen kompetenzorientierten Ansatz. Das nationale Curriculum definiert sogenannte *Key Competencies*,

darunter *Thinking*, *Relating to Others*, *Managing Self* sowie *Participating and Contributing* (New Zealand Ministry of Education 2007). Diese Kompetenzen werden überwiegend im Kontext fachlicher Lernprozesse beobachtet und bewertet. Die Leistungsbewertung ist in Neuseeland stark von *Assessment for Learning* geprägt: Lehrkräfte arbeiten mit Portfolios, projektbasierten Aufgaben und dialogischen Feedbackprozessen (New Zealand Qualifications Authority, o. J.). In der Sekundarstufe II nutzt Neuseeland mit dem *National Certificate of Educational Achievement* (NCEA) ein modulares Prüfungssystem. Leistungen werden dort über sogenannte *Achievement Standards* nachgewiesen, die sowohl interne Leistungsnachweise wie Projekte oder Präsentationen als auch externe Prüfungen umfassen. Forschungen zeigen jedoch, dass diese Flexibilität auch Probleme erzeugt: Lehrkräfte müssen aktiv entscheiden, welche Standards sie für ihre Kurse auswählen, und werden dadurch faktisch zu Curriculum-Designer:innen (Hipkins & Cameron 2018). Zugleich besteht die Gefahr, bevorzugt solche Standards zu wählen, die für Schüler:innen leichter zu erfüllen sind, was zu einer Fragmentierung des Curriculums führen kann. Die Forschung betont daher, dass Curriculum, Unterricht und Assessment kohärent miteinander verbunden werden müssen, damit die im Curriculum verankerten *Key Competencies* tatsächlich sichtbar und bewertbar werden. Neuseeländische Studien zeigen zudem, dass viele *Assessment for Learning*-Praktiken situativ im Unterricht entstehen und deshalb nur begrenzt systematisch überprüfbar sind; sie können wertvolle Informationen für Lernförderung und Unterrichtsentscheidungen liefern, eignen sich jedoch weniger für Standardisierung und Vergleichbarkeit (ebd.).

5.6 Singapur

> **Singapur hat die halbjährlichen Zwischenprüfungen abgeschafft und fördert im damaligen Zeitraum der Prüfungsvorbereitung nun *21st Century Skills*.**

Singapur gilt als Pionier bei der systematischen Vermittlung von *21st Century Competencies* (21CC) im Schulsystem. Bereits 2010 führte das Bildungsministerium einen umfassenden 21CC-Rahmen ein, der neben Fachwissen auch kritisches und kreatives Denken, Kommunikation, Zusammenarbeit, Selbstregulation und *Global Citizenship* umfasst. Diese Kompetenzen sollen systematisch in der schulischen Bildung verankert werden (Singapore Ministry of Education 2024a). In der Umsetzung fokussierte Singapur zunächst vor allem auf veränderte Bildungspläne und innovative Pädagogik, die projektbasierte Lernphasen, *Inquiry-Based Learning* und außerschulisches Lernen durch Engagement (*Service Learning*) im Gemeinwesen vorsieht. Im Bereich der Leistungsbewertung wurden zunächst vor allem in der sechsjährigen Grundschule und der Sekundarstufe I Veränderungen vorgenommen: An Grundschulen wurden standardisierte Prüfungen

in den ersten Schuljahren abgeschafft oder reduziert, um Raum für formatives Feedback und ganzheitliche Prüfungsformate zu schaffen. Jüngere Schüler:innen erhalten statt Noten ausführliche Rückmeldungen zu ihren Lernfortschritten, einschließlich der Entwicklung ihrer sozialen und methodischen Kompetenzen. In der Sekundarstufe I arbeiten Schulen zudem mit digitalen Portfolios und Reflexionsjournals. Darüber hinaus werden ab 2028 für den Zugang zu Junior Colleges (weiterführenden Schulen mit Oberstufe) weniger Prüfungsfächer berücksichtigt; seit 2026 sind außerdem die halbjährlichen Zwischenprüfungen abgeschafft. Die bislang vorgesehene dreiwöchige Vorbereitungszeit wird nun genutzt, um gezielt 21st Century Skills jenseits der in Asien traditionell dominanten High-Stakes-Prüfungen zu fördern (Lam et al. 2025). In diesem Transformationsprozess wurde deutlich, dass Eltern durch Erwartungen und Praktiken wesentlich dazu beitragen, traditionelle, stark leistungs- und prüfungsorientierte Erfolgsvorstellungen im System aufrechtzuerhalten und somit eine ausgeprägte Testkultur stabilisieren (Hardy et al. 2021). Singapur reagierte darauf mit systematischer Einbindung und transparenter Kommunikation, etwa durch Initiativen zur Rolle der Eltern, die als zentrale Stellschraube identifiziert wurden, um Prüfungsstress der Schüler:innen zu reduzieren und die mentale Gesundheit zu stärken (Singapore Ministry of Education 2024b).

Im Curriculum Singapurs wurden zudem jahrzehntelang bestehende Selektionsmechanismen reformiert. Mit der Einführung fachbezogener Leistungsdifferenzierung können Schüler:innen einzelne Fächer auf unterschiedlichen Niveaus belegen, statt einem festen Bildungsgang zugeordnet zu werden. Dadurch soll mehr Flexibilität entstehen und die frühere Stigmatisierung leistungsschwächerer Gruppen reduziert werden. Gleichzeitig bleibt die Zuweisung zu weiterführenden Schulen eng an die Ergebnisse zentraler Abschlussprüfungen gebunden. Deutliche Unterschiede zwischen besonders leistungsstarken und weniger prestigeträchtigen Schulen bestehen daher fort. Tan (2024) zeigt, dass die Reform zwar mehr Durchlässigkeit innerhalb von Schulen schafft, die leistungsbasierte Trennung zwischen Schulen aber bestehen lässt. Das System bewegt sich damit weiterhin im Spannungsfeld von Chancengerechtigkeit und Selektionslogik.

Auf der Ebene der zentralen Prüfungen – insbesondere PSLE sowie GCE O- und A-Levels – ist Singapur traditionell stark leistungsorientiert und arbeitet mit inhaltsbezogenen Tests. Gleichwohl wurde auch hier Raum für komplexere Aufgaben geschaffen. Ein markantes Beispiel ist das Pflichtfach Project Work im Junior College, das in den frühen 2000er-Jahren eingeführt wurde: Schüler:innen bearbeiten in Gruppen ein mehrmonatiges fächerübergreifendes Projekt und erstellen einen Bericht sowie eine Präsentation. Bewertet werden sowohl das Gruppenprodukt als auch individuelle Leistungen, unter anderem in einer mündlichen Prüfung. *Project Work* erscheint mit einer eigenen Note im A-Level-Zertifikat und soll Fähigkeiten wie kollaboratives Problemlösen, selbstständiges Forschen, Kreativität und Kommunikationskompetenz fördern und prüfen.

Empirische Untersuchungen aus Singapur zeigen, dass Schüler:innen durch *Project Work* beispielsweise ihre Team- und Präsentationsfähigkeiten signifikant verbessern und die damit verbundene Prüfungsleistung als

sinnvolle Ergänzung zu schriftlichen Prüfungen wahrnehmen (Tan et al. 2017). Eine aktuelle Studie in der Tradition von Tan und Kolleg:innen verweist zudem auf positive Effekte computergestützter Verfahren im Sinne von *Learning Analytics* auf komplexe Kompetenzen (Koh et al. 2023). Diese Effekte hängen allerdings stark von der didaktischen Einbettung ab: In einer Schule mit expliziter Einführung und kontinuierlicher Begleitung nahmen selbstberichtete *Deeper-Learning*-Kompetenzen zu, ebenso die wahrgenommenen kritischen Lesefähigkeiten zur Analyse und Einordnung von (Des-)Informationen. Ohne eine solche Rahmung blieben die Effekte aus oder kehrten sich teilweise sogar ins Negative um.

Singapur hat zudem wertvolle Erfahrungen mit der Skalierung kollaborativer Kompetenzen in institutionellen Settings gesammelt. Neuere Forschung zeigt, dass Innovationen nicht allein durch Pilotprojekte tragfähig werden, sondern einer systematischen organisationalen Verankerung bedürfen. Suat Hoon und Kolleg:innen (2025) untersuchen die Skalierung eines institutionellen *Teamwork Assessment* und zeigen, dass hohe Akzeptanz insbesondere dann entsteht, wenn strukturelle Rahmenbedingungen (klare *Policies*, Ressourcen, digitale Infrastrukturen und strategische Unterstützung durch das Management) vorhanden sind. Die institutionelle Verankerung vollzieht sich dabei in einem vierstufigen Prozess: *Co-Experience*, *Co-Definition*, *Co-Elevation* und *Co-Development*.

Mit dem Singapore Student Learning Space (SLS) verfügt Singapur seit 2018 über eine landesweit implementierte digitale Lern- und Assessmentplattform, die allen öffentlichen Schulen zur Verfügung steht. SLS fungiert nicht nur als cloudbasiertes Speichersystem, sondern als integrierte Infrastruktur für formative Leistungsrückmeldung, adaptive Lernpfade und seit einem Update im Oktober 2025 KI-gestützte Feedbackprozesse (Singapore Student Learning Space 2025). Lehrkräfte können offene Aufgabenformate mit *Rubrics* verknüpfen, automatische Kurzantwort-Feedbacksysteme nutzen und adaptive Diagnoseinstrumente einsetzen. Lernende erhalten personalisierte Rückmeldungen, dokumentieren Lernprozesse in digitalen Notebooks und greifen auf lernzielbezogene Empfehlungen zurück. Auf diese Weise institutionalisiert Singapur formative *Learning Analytics* und schafft eine professionelle datenbasierte Grundlage, auf der summative Entscheidungen angereichert werden können (Singapore Ministry of Education 2025).

Um die Beurteilung solcher komplexeren Prüfungsformate zu professionalisieren, erhalten Lehrkräfte Training im *Authentic Assessment*. Zudem wird Wert darauf gelegt, dass Prüfungen auf höhere Anforderungsbereiche ausgerichtet werden durch Anteile von Anwendung, Analyse und Evaluation in den schriftlichen Examensaufgaben (vgl. sogenannte *Higher-Order-Thinking*-Fragen in Mathematik und Naturwissenschaften). Das Land erzielt in internationalen Vergleichen (PISA, TIMSS) weiterhin herausragende Ergebnisse und gleichzeitig fällt auf, dass Schüler:innen aus Singapur in domänenübergreifenden Kompetenzen (z. B. PISA: *Creative Problem Solving* und *Collaborative Problem Solving* seit 2015) ebenfalls hervorragend abschneiden.

Die anhaltend sehr guten Leistungen Singapurs in internationalen Vergleichen, auch in PISA und TIMSS, werden zum Teil darauf zurückgeführt, dass die Bildungspolitik frühzeitig einen ganzheitlicheren Kompetenzbegriff eingeführt hat (Tan et al. 2017). Zugleich diskutiert Singapur verstärkt den Einsatz simulationsbasierter Computertests und adaptiver Verfahren, etwa WiREAD+ zur Förderung kritischer Lesekompetenz oder *My Groupwork Buddy* zur Entwicklung und Reflexion von Teamkompetenz. Im Rahmen des *EdTech Masterplan 2030* integriert das Bildungsministerium KI systematisch in Lern- und Bewertungsprozesse. KI-gestützte Lernanalysen, adaptive Systeme und automatisierte Feedbackmechanismen sollen personalisiertes Lernen ermöglichen und diagnostische Prozesse verbessern (Singapore Ministry of Education 2026). Gleichzeitig wird die Entwicklung von *AI Literacy* selbst als Bestandteil schulischer Kompetenzbildung verstanden. Trotz der deutlichen Tendenz zu *Assessment for Learning* bleibt die Selektionswirkung von Prüfungen bestehen; Bildungsforschung und Bildungspolitik versuchen daher, *High-Stakes*-Prüfungen und ganzheitliche Kompetenzentwicklung nebeneinander zu ermöglichen und die Reformeffekte kontinuierlich zu evaluieren (Hardy et al. 2021; Siow Huang 2025; Wong et al. 2020).

Dafür hat Singapur institutionelle Forschungsstrukturen zur systematischen Weiterentwicklung von Assessment aufgebaut. Am National Institute of Education arbeiten unter anderem das Centre for Assessment and Evaluation sowie das Centre for Research in Pedagogy and Practice daran, evidenzbasierte Erkenntnisse zu Lehr-Lern-Assessment zu generieren und direkt in Bildungspolitik und Lehrer:innenbildung zu übertragen. Ziel ist es, *Assessment Literacy* bei Lehrkräften zu stärken und Reformen wissenschaftlich zu begleiten.

Insgesamt bietet Singapur ein instruktives Beispiel dafür, wie ein traditionell prüfungsstarkes System schrittweise, sehr bewusst und entschlossen, ohne erkennbare Absenkung seiner Standards innovative Prüfungsformate integriert.

5.7 International Baccalaureate (IB)

Das *International Baccalaureate (IB)* gilt international als leistungsstarkes und innovationsorientiertes Bildungssystem, das konsequent auf die Förderung komplexer Kompetenzen ausgerichtet ist. Dieser Abschluss mit weltweiter Hochschulzugangsberechtigung wird von Genf aus gesteuert und an internationalen Schulen weltweit angeboten. Forschungen zu seinen curricularen und prüfungskulturellen Arrangements liefern daher aufschlussreiche Hinweise für die Weiterentwicklung zeitgemäßer Prüfungsformate. Zugleich zeigt sich bislang, dass komplexe Kompetenzen auch im IB in Prüfungen eher unterbelichtet bleiben. Insbesondere reflexive, kontextgebundene und schwer standardisierbare Kompetenzen werden eher indirekt und häufig über qualitative Lernsituationen jenseits klassischer Prüfungen adressiert (Shukur 2025).

Eine internationale Evaluationsstudie des Oxford University Centre for Educational Assessment zeigt zudem, dass Schüler:innen im *IB Middle Years Programme* signifikant höhere Werte im kritischen Denken erreichen als Vergleichsgruppen. Die Studie führt diese Effekte auf das Zusammenspiel von kriterialer Bewertung, kontinuierlichem Feedback und lernprozessbegleitenden Assessmentpraktiken zurück: „Assessment is an instrumental element to support and encourage students' critical thinking development" (McGrane et al. 2022, S. 19). Gleichzeitig wird deutlich, dass selbst in einem Hochleistungsschulsystem wie dem IB kriteriale Bewertungsraster für komplexe Kompetenzen – etwa *Critical Thinking* – ein *Teaching to the Test* erzeugen können, wenn *Rubrics* als abzuarbeitende Checklisten statt als Instrumente für komplexe Lernprozesse genutzt werden (ebd., S. 54).

Mit den seit 2023/2024 laufenden Pilotierungen rund um den IB-Kurs *Systems Transformation* erprobt das IB eine neue Logik von Curriculum und Leistungsbewertung: weg von primär fachdisziplinär gerahmten, stark auf die Abschlussprüfung hin ausgerichteten Prüfungsformaten, hin zu einem stärker projekt- und kompetenzorientierten Arrangement, das *Real-World Learning*, *Authentic Assessment* und eine größere Autonomie der Schüler:innen ausdrücklich als Entwicklungsziele benennt (International Baccalaureate 2026b; UWC South East Asia, o. J.).

Im Zentrum steht ein doppeltes Pilot-Setting: erstens der Kurs *IB Systems Transformation: Leadership for Change* am United World College of South East Asia (UWCSEA) in Singapur, der seit August 2024 läuft, und zweitens die *Systems Transformation Pathway: Leadership for Just Futures* am UWC Atlantic College in Wales, die seit 2023 erprobt wird. Beide Formate sind im Rahmen der kollaborativen 16+-Überprüfung des IB als *Alternative Assessment Pathway* ausgewiesen, mit dem perspektivischen Ziel, nach Auswertung der Pilot- und Evaluationsdaten eine weltweite Skalierung auf weitere IB-Schulen zu ermöglichen (International Baccalaureate 2023, 2024, 2026b, 2026c; UWC South East Asia, o. J.).

Singapur und Wales als zwei Varianten derselben Reformidee

Der Kurs am UWCSEA ist als *Double Standard Level Course* mit insgesamt 300 Unterrichtsstunden konzipiert und ersetzt damit zwei *Standard-Level*-Fächer; zugleich wird er in der IB-Logik als *Dual Award* geführt, das heißt, es werden zwei identische Noten ausgewiesen (International Baccalaureate 2024). Diese curricularen Vorgaben sind bewusst nicht als klassische Disziplinen angelegt, sondern als transferorientierte Kompetenz- und Handlungsfelder – darunter *Systems Thinking*, *Design Thinking*, *Project* beziehungsweise *Change Management*, *Leadership* und *Changemaker Mindset* –, die über zwei Jahre hinweg im Modus projektbasierten Lernens aufgebaut und nachgewiesen werden sollen (UWC South East Asia, o. J.).

Die Wales-Variante am UWC Atlantic ist stärker als *Pathway* gerahmt: Sie verbindet drei *Higher-Level*- und ein *Standard-Level*-Fach mit einem Kurs, der auf vier *Impact Areas* fokussiert – *Food*, *Biodiversity*, *Energy* und *Migration* – und Transformation als Bearbeitung komplexer Gerechtigkeits- und Nachhaltigkeitsfragen versteht (International Baccalaureate 2023). Während das UWCSEA-Design seine Legitimation stärker über generische, berufs- und zukunftsbezogene Transferkompetenzen bezieht, rahmt UWC

Atlantic die Idee stärker über ökologische und soziale Gerechtigkeitsproblemlagen. Zugleich verweist die Schule bereits auf eine Pilotkohorte von 2023 bis 2025 und eine angestrebte Ausweitung nach Evaluation (International Baccalaureate 2023).

Assessment-Design: *Authentic Assessment* als strukturprägendes Prinzip
Der vielleicht innovativste Kern der Pilotierung liegt im Assessment. In beiden Settings wird ein vierteiliges Modell implementiert, das externe und interne Komponenten kombiniert und zentrale Lernprodukte – Portfolio- beziehungsweise Artefaktarbeit, Fallanalyse, Projektplanung und Impact-Analyse – als Beweisformen von Kompetenzentwicklung nutzt (International Baccalaureate 2026b; UWC South East Asia, o. J.). Dadurch verschiebt sich die Logik der Leistungsbewertung von einer punktuellen Abschlussprüfung hin zu einem über zwei Jahre verteilten, artefakt- und performanzbasierten Prüfungsarrangement. Explizit wird diese Verschiebung auch als Abkehr von einer reinen Abschlussprüfung hin zu einer Portfoliologik beschrieben (UWC South East Asia, o. J.).

UWCSEA (Singapur): Die öffentlich zugängliche UWCSEA-Dokumentation beschreibt vier *coursework*-basierte Komponenten: extern eine *Impact Analysis* (30 %) und ein *Competencies Portfolio* (20 %), intern eine *Comparative Case Study* (20 %) und einen *Project Proposal* (30 %). Assessment ist dabei ausdrücklich über die zwei Kursjahre verteilt und nicht als traditionelle Abschlussklausur organisiert (UWC South East Asia, o. J.).

UWC Atlantic (Wales): Öffentlich zugängliche IB-Informationen bestätigen auch das UWC Atlantic College als Pilotschule der *Systems-Transformation*-Entwicklung in Wales. Anders als für UWCSEA sind die öffentlich zugänglichen Angaben zu den konkreten Gewichtungen der Assessment-Komponenten jedoch deutlich knapper; belastbar nachweisbar sind vor allem die Pilotierung selbst, die Verankerung im 16+-Review sowie die Erprobung neuer Evidenz- und Assessmentformen (International Baccalaureate 2023, 2026b, 2026c).

Über beide Varianten hinweg lassen sich drei Verschiebungen erkennen: Erstens wird versucht, die Kompetenzentwicklung über Artefakte (kuratierte Portfolioelemente, dokumentierte Projektarbeit) prüfbar zu machen, wodurch Lernprozesse stärker sichtbar werden als in reinen Produktexamina (International Baccalaureate 2026b; UWC South East Asia, o. J.). Zweitens wird Authentizität nicht nur als pädagogisches Postulat, sondern als Konstruktionsprinzip der Leistungsbewertung implementiert (UWC South East Asia, o. J.). Drittens wird durch die Kombination aus externer Bewertung und moderierter interner Bewertung eine Hybridlogik von Standardisierung und Kontextsensitivität erprobt, um den bereits benannten Zielkonflikten von Prüfungskulturen entgegenzuwirken.

Für die Frage nach der IB-Prüfung ist entscheidend, dass *Leadership for Change* in der Darstellung des UWCSEA explizit als vollständig *coursework*-basiert beschrieben wird – also als Format, das „assessed entirely by coursework … rather than the traditional end of course exam“ ist (UWC South East Asia, o. J.). Das bedeutet nicht, dass es keine prüfungsförmigen Situationen mehr gibt; die *Case-Study*-Komponente arbeitet etwa mit einem

vom IB produzierten Fall und verbindet kollaborative sowie individuelle Leistungsanteile (ebd.). Die Gesamtlogik verlagert den Schwerpunkt jedoch deutlich von einer einmaligen, hochgewichteten Abschlussprüfung hin zu einer über die Kursdauer verteilten, mehrperspektivischen Evidenzsammlung. Genau darin liegt der Reformanspruch des *Alternative Assessment Pathway* im Rahmen der 16+-Überprüfung (International Baccalaureate 2026c).

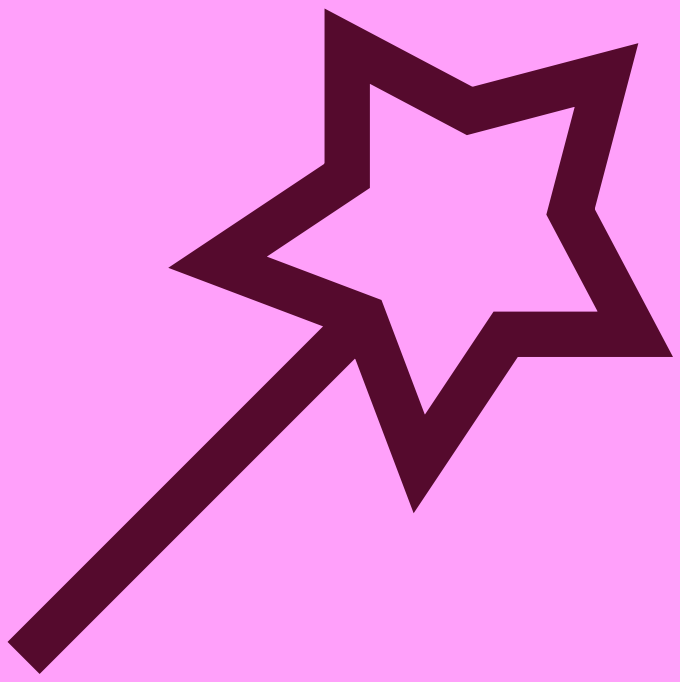
Die Pilotierung fällt in eine Phase, in der generative KI die Assessment-Debatte mit hoher Geschwindigkeit verändert. Das IB positioniert sich dabei doppelt: Einerseits betont es, dass KI-Tools Lernprozesse unterstützen können; andererseits müssen KI-generierte Beiträge transparent gemacht und korrekt referenziert werden (International Baccalaureate 2025). Zugleich hält das IB an menschlicher Bewertung fest und beschreibt KI eher als mögliches Qualitätssicherungsinstrument im Bewertungsprozess denn als Ersatz für die Urteilsfähigkeit menschlicher Prüfer:innen (ebd.).

Mit Blick auf die Pilotkurse ist daran anschlussfähig, dass portfolio- und projektbasierte Leistungen nicht automatisch „KI-resistent“ sind, aber andere Integritätslogiken erlauben als klassische Klausuren: Prozessdokumentation, Artefaktgenese, Feedbackschleifen und die Nachvollziehbarkeit von Entscheidungen können zu prüfungsrelevanten Ankerpunkten werden. Normativ gerahmt wird dies durch die *IB Academic Integrity Policy*, die (in der aktualisierten Fassung) explizite Richtlinien zur Nutzung von KI-Tools integriert (International Baccalaureate 2026a). Ergänzend diskutiert das IB öffentlich, ob KI eine „Krise“ oder eine „Chance“ für Assessment darstellt und hebt dabei unter anderem die authentischen Aufgabenformate hervor, die Verständnis, Urteilskraft und Transfer stärker einfordern (International Baccalaureate 2025).

Die Bedeutung der Pilotierungen in Singapur und Wales liegt daher weniger in der Einführung eines neuen Schulfaches als in der Erprobung eines programmatischen Assessment-Paradigmas, das für die Weiterentwicklung von Abschlussprüfungen insgesamt richtungsweisend sein kann: Erstens wird Kompetenzorientierung nicht nur curricular gefordert, sondern über evidenzbasierte Nachweisformen operationalisiert. Zweitens werden Kollaboration und projektförmige Problembearbeitung als prüfungsrelevante Praktiken gerahmt. Drittens werden externe und schulbasierte Bewertung so kombiniert, dass sowohl Vergleichbarkeit als auch Kontext- und Prozesssensitivität möglich werden. Viertens richtet sich Assessment in einer KI-geprägten Welt stärker auf dokumentierte Urteils- und Gestaltungsleistungen, die sich nicht auf bloße Textproduktion reduzieren lassen (International Baccalaureate 2025, 2026b; UWC South East Asia, o. J.).

Bemerkenswert ist schließlich, dass das IB diese Erprobung ausdrücklich als mehrjährige Programmentwicklung beschreibt, eben mit den klar benannten Pilotorten in Wales und Singapur und einem potenziellen langfristigen Ziel einer globalen Einführung ab 2030 (International Baccalaureate 2026b, 2026c). Assessment-Reform erscheint hier nicht als punktuelle Revision einzelner Prüfungsformate, sondern als langfristig angelegte Programmentwicklung, die Fragen komplexer Bildungsziele,

veränderter Lern- und Prüfungskulturen sowie der – auch juristischen – Legitimation von Leistungsnachweisen für hochschulzugangsrelevante Abschlussprüfungen zusammenführt. Gerade darin liegt die eigentliche Neuartigkeit der aktuellen IB-Entwicklung und ihr Inspirationspotenzial für die Diskussion über die Weiterentwicklung des Abiturs in Deutschland.



6
Potenziale von Künstlicher Intelligenz (KI) in der Prüfungskultur

Die disruptive Kraft Künstlicher Intelligenz (KI) prägt zunehmend auch den Diskurs um Prüfungskulturen. Während Fragen des Datenschutzes, der Rechenschaftspflicht und der Verantwortungszuschreibung weiterhin zentral bleiben, fokussiert dieser Beitrag vor allem die Qualität KI-gestützter Messverfahren. Einen ersten umfassenden Orientierungsrahmen für diese Debatte haben die Europäische Kommission und die OECD (2026) mit dem *Framework Empowering Learners for the Age of AI* vorgelegt, das die Implementation von KI systematisch mit Fragen der Kompetenzentwicklung und des Assessments verknüpft.

6.1 Von computergestütztem Testing zu KI-basiertem Assessment

Bisherige computergestützte Testverfahren zeichneten sich vor allem durch hohe Reliabilität und Standardisierbarkeit aus, waren jedoch nur begrenzt in der Lage, komplexe und kontextgebundene Kompetenzdimensionen zu erfassen. KI-basierte Systeme, insbesondere *Large Language Models* (LLMs), verschieben dieses Verhältnis: Sie können offene, vielschichtige Datenformate verarbeiten und qualitativ differenziertes Feedback generieren. Damit entstehen neue automatisierte Möglichkeiten, auch komplexere Kompetenzkonstrukte datenbasiert zu erfassen. Studien wie die von Bulut und Beiting-Parrish (2024) systematisieren die technologischen Potenziale von KI im Assessment entlang mehrerer Dimensionen: automatisierte Bewertung offener Antworten, schnelle Rückmeldungen mittels *Natural Language Processing, Learning Analytics* zur Diagnose von Lernständen sowie die Vorhersage von Lernverläufen oder Abbruchrisiken. Assessment wird dadurch skalierbarer, schneller und stärker datengetrieben. Gleichzeitig entstehen neue Prüfungsformen. Diskutiert werden automatisierte Item-Erstellung, die Generierung von Aufgaben, die Bewertung offener Formate, Kontrollsysteme zur Sicherung von Prüfungsintegrität sowie adaptive Feedbacksysteme. Vor allem für komplexe performanzbasierte Aufgaben erweitert sich dadurch das Repertoire erheblich.

Eine besondere Stärke KI-gestützter Systeme liegt in der Auswertung von Prozessdaten. In komplexen Aufgabenformaten fallen zahlreiche Interaktionsdaten an, die für menschliche Beurteilende nur begrenzt zugänglich sind. KI kann diese Daten analysieren und Muster identifizieren, etwa Bearbeitungsstrategien, Zeitverläufe oder Formen der Interaktion in kollaborativen Settings. Arbeiten von Woolley und Kolleg:innen (2021) zeigen beispielsweise, dass sich Rollenverteilungen und Interaktionsqualitäten in Gruppenprozessen datenbasiert rekonstruieren lassen. Dazu eröffnen sich vielversprechende Potenziale für die individuelle Förderung selbstregulierten Lernens mit generativer KI (Joarder & Chatti 2026).

Darüber hinaus ermöglicht KI adaptive Prüfungsformate, die Aufgaben dynamisch an den Kompetenzstand der Lernenden anpassen. Solche adaptiven Systeme können differenziertere Leistungsfeststellungen ermöglichen und eröffnen auch im Bereich des Aufgabendesigns neue Möglichkeiten.

6.2 Grenzen und Validitätsprobleme KI-gestützter Bewertung

Trotz dieser Potenziale stellen sich grundlegende messmethodische Herausforderungen. Generative KI basiert auf probabilistischen, modellbasierten Vorhersagen. Ihre Outputs sind daher nicht als Wahrheitsurteile, sondern als Wahrscheinlichkeitsaussagen zu verstehen (Runge et al. 2024).

> **Ähnliche Urteile der KI bedeuten nicht, dass die zugrunde liegenden Bewertungslogiken identisch sind.**

Empirische Studien zeigen zwar häufig eine hohe Übereinstimmung zwischen KI-basierten Bewertungen und menschlichen Ratings. Lorié und Dadey (2026) weisen jedoch darauf hin, dass diese Übereinstimmung nicht zwingend auf derselben Konstruktbasis beruht. Eine ähnliche Bewertung bedeutet also noch nicht, dass dieselbe Kompetenzdimension reliabel erfasst wurde. KI kann zu ähnlichen Urteilen gelangen, ohne dass die zugrunde liegenden Bewertungslogiken identisch sind.

Hinzu kommen Probleme der Stabilität und Reproduzierbarkeit. Hao und Kolleg:innen (2024) zeigen, dass identische Antworten je nach Eingabe-Prompts oder Modellkonfiguration unterschiedlich bewertet werden können. Dadurch wird die Nachvollziehbarkeit von Bewertungsentscheidungen eingeschränkt, was ein gravierendes Problem insbesondere in *High-Stakes*-Prüfungen darstellt. Zugleich reproduzieren KI-Systeme die in ihren Trainingsdaten enthaltenen Muster und Bewertungslogiken, wodurch bestehende Bias-Strukturen fortgeschrieben werden können. Das stellt eine zentrale Herausforderung für faire und gerechte Leistungsbewertung dar.

> ***Metric Substitution*: Die Verengung auf jene Kompetenzdimensionen, die in Prüfungen vergleichsweise leicht messbar sind.**

Vor diesem Hintergrund lässt sich die Validität KI-gestützter Bewertungssysteme nicht allein durch technische Optimierung sichern. Die Validitätsfrage verschiebt sich vielmehr auf die Ebene der Einbettung solcher Systeme in konkrete Prüfungsprozesse. Barthakur und Kolleg:innen (2026) warnen in diesem Zusammenhang vor *Metric Substitution*, also einer Verengung auf jene Kompetenzdimensionen, die in Assessment-Situationen vergleichsweise leicht messbar sind. Mit der wachsenden Verbreitung generativer KI und datenbasierter Tools gewinnt daher *Assessment Literacy* zusätzlich an Bedeutung, damit Lehrkräfte die Potenziale und Grenzen dieser Messverfahren professionell einordnen können.

Aktuelle Policy-Empfehlungen betonen deshalb, dass KI-basierte Bewertungen systematisch gegen externe Kriterien und menschliche Urteile validiert werden müssen. Zudem erfordert ihr Einsatz standardisierte Rahmenbedingungen wie eine kontrollierte Promptgestaltung, transparente Modellkonfiguration und kontinuierliches Monitoring der Bewertungskonsistenz (Lorié & Dadey 2026). Zentral ist dabei das Prinzip des *Human-in-the-Loop*: KI fungiert nicht als autonomes Bewertungssystem, sondern als Bestandteil eines ko-konstruktiven Bewertungsprozesses, in dem menschliche Akteur:innen die Verantwortung für Interpretation und Entscheidung behalten. Implementationsformen bewegen sich auf einem Kontinuum zwischen Unterstützung und Substitution; gerade in selektionsrelevanten Kontexten sind daher klare und auch juristisch geklärte Grenzen KI-gestützter Beteiligung erforderlich.

6.3 Systemische Implikationen für Prüfungskulturen

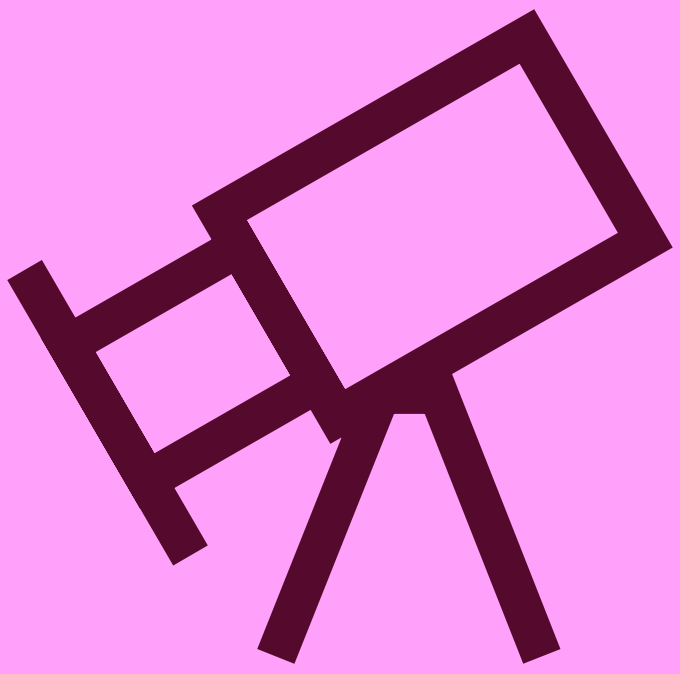
Die Integration von KI stellt bestehende Prüfungsformate grundsätzlich infrage. Beispiele wie Dänemark, das bestimmte Abschlussprüfungen im Zuge der Verfügbarkeit generativer KI zurückziehen beziehungsweise neu ordnen musste, verdeutlichen den Anpassungsdruck selbst auf fortschrittliche Prüfungssysteme.

Internationale Organisationen wie die UNESCO argumentieren deshalb, dass Bildungssysteme Prüfungsformate aktiv weiterentwickeln müssen, statt am Status quo festzuhalten (UNESCO 2023). KI wird langfristig integraler Bestandteil von Lern- und Arbeitsprozessen sein; damit verschieben sich auch die Anforderungen an Prüfungen. Für eine nationale KI-Strategie lohnt sich in diesem Zusammenhang ein erneuter Blick nach Estland, wo im März 2025 die landesweite Initiative *AI Leap* gestartet ist.

AI Leap: Eine auf ChatGPT basierende Lernanwendung liefert nicht einfach fertige Antworten, sondern stößt durch gezielte Rückfragen Denkprozesse an.

Dort wird KI landesweit in der Sekundarstufe II eingeführt und konsequent pädagogisch gerahmt. Ausgangspunkt ist die Beobachtung, dass ein großer Teil der Lernenden bereits informell generative KI nutzt, dies aber häufig in einer Weise geschieht, die eher auf Effizienzsteigerung als auf Kompetenzentwicklung zielt. *AI Leap* setzt genau hier an: Eine eigens entwickelte, auf ChatGPT basierende Lernanwendung liefert nicht einfach fertige Antworten, sondern stößt durch gezielte Rückfragen Denkprozesse an und unterstützt metakognitive Strategien.

Auffällig ist die systematische Einbettung auf Landes- und Schulebene sowie ihre transparente Evaluation inklusive messbarer Ziele: Inzwischen bestehen in über 80 Prozent der Schulen KI-Lerngemeinschaften, in denen Lehrkräfte den Einsatz von KI gemeinsam reflektieren und weiterentwickeln, gleichzeitig bleibt die App-Nutzung der Schüler:innen mit 20 Prozent wöchentlicher Aktivität deutlich unter den Erwartungen. KI wird im Rahmen dieser estnischen Strategie nicht als individuelles Tool, sondern als Bestandteil professioneller Unterrichtsentwicklung verstanden. Im Vordergrund steht weniger die punktuelle Prüfung als die fortlaufende Kompetenzdiagnostik. Erste Ansätze deuten darauf hin, dass KI genutzt wird, um individuelle Lernpfade zu analysieren, adaptive Aufgaben bereitzustellen und Lehrkräften evidenzbasierte Hinweise zur Unterstützung von Lernenden zu geben. Bemerkenswert ist schließlich die konsequente wissenschaftliche Begleitung. In Kooperation mit der Universität Tartu, der Universität Stanford und OpenAI wird eine groß angelegte Wirkungsstudie durchgeführt, die untersucht, unter welchen Bedingungen KI-gestützte Lernumgebungen tatsächlich zur Kompetenzentwicklung beitragen. Eine Initiative, die in den kommenden Monaten aufmerksam zu beobachten sein wird.



7
Handlungsperspektiven für Deutschland

Aus den internationalen Entwicklungen, den diskutierten Praxisbeispielen sowie der damit einhergehenden Forschung lassen sich verschiedene Handlungsperspektiven ableiten, die deutsche Prüfungskulturen stärker an die Entwicklungsrichtung aktueller internationaler Hochleistungsschulsysteme anschließen könnten. Die folgenden Vorschläge verstehen sich dabei als mögliche politische und institutionelle Ansatzpunkte für die Weiterentwicklung der gymnasialen Oberstufe in Deutschland.

7.1 Rechtliche Rahmen für formative und kombinierte Leistungsnachweise erweitern

Eine moderne Prüfungskultur verbindet Leistungsnachweise mit kontinuierlicher Lernbegleitung. Vor diesem Hintergrund sollten komplexe Leistungsnachweise nicht primär als punktuelle und optionale Ersatzformate für Klausuren verstanden werden, sondern als Prüfungsarchitekturen, die Lernprozesse kontinuierlich sichtbar machen. Entscheidend ist, gerade im Zeitalter von Künstlicher Intelligenz, dass Schüler:innen nicht nur Endprodukte vorlegen, die bewertet werden, sondern ihr Vorgehen reflektieren, Entscheidungen begründen und fachliches Wissen in authentischen Situationen anwenden und dabei begleitet werden.

Bundesländer sollten dafür in ihren Oberstufenverordnungen einen klar benannten, aber offen gestalteten Typus komplexer Leistungsnachweise verankern. Diese Nachweise könnten produktive, reflexive, schriftliche, mündliche, individuelle und gegebenenfalls kollaborative Elemente verbinden. Ihre Stärke liegt gerade darin, dass sie unterschiedliche Evidenzen für Kompetenzentwicklung zusammenführen und dadurch mehr sichtbar machen als traditionelle Klausuren.

Die Länder sollten diese Formate nicht durch kleinteilige Vorgaben verengen, sondern Qualitätsrahmen entwickeln. Sinnvoll wären landesweite Handreichungen mit Kompetenzdefinitionen, exemplarischen Aufgaben, Bewertungsbeispielen und Mindestanforderungen zu transparenter Aufgabenstellung, kriterialer Bewertung, Dokumentation, fairer Hilfsmittelnutzung und individueller Zurechenbarkeit. Auf diese Weise könnten formative Leistungsnachweise rechtssicher verankert werden, ohne die Kompetenzdimensionen durch Prüfungsvorgaben zu verengen.

7.2 Selektionsärmere Entwicklungsräume und curriculare Freiräume schaffen

International zeigt sich zunehmend der Trend, stark selektionsorientierte Lern- und Prüfungskontexte zugunsten projektorientierter, kollaborativer und reflexiver Arbeitsformen zu reduzieren. In Anbetracht der vergleichsweise starren deutschen Abiturvorgaben erscheint insbesondere die Einführungsphase als zentraler Entwicklungsraum. Sie ist weniger stark durch Abschlussprüfungen geprägt und bereitet zugleich auf die Qualifikationsphase vor. Das irische Modell des *Transition Year* könnte hierfür als Orientierung dienen, weil es Schüler:innen Raum für persönliche Schwerpunktsetzung, projektorientiertes Lernen und reflektierte Bildungsentscheidungen eröffnet.

Damit solche Entwicklungsräume wirksam werden können, müssen Prüfungsreformen mit Curriculums- und Unterrichtsentwicklung zusammengedacht werden. Maßgeblich ist dabei das Prinzip des *Constructive Alignment*: Lernziele, Lernaktivitäten und Prüfungsformate müssen kohärent aufeinander bezogen sein. Länder sollten Curricula daher stärker auf fachliche Schlüsselkompetenzen, exemplarische Vertiefung und komplexe Anwendungssituationen ausrichten und diese mit überfachlichen Kompetenzzielen verbinden. Orientierung bieten etwa Curricula wie das australische *F–10 Curriculum*, in dem fachliche und überfachliche Kompetenzen systematisch aufeinander bezogen werden (Australian Curriculum, Assessment and Reporting Authority 2022).

Eine mögliche Umsetzung bestünde darin, projektorientierte Leistungsnachweise in der Einführungsphase zu stärken, Portfolioarbeit auszubauen, niedrigschwellige oder unbenotete Kompetenznachweise zu erproben und Schüler:innen bei der individuellen Profilbildung stärker zu beraten. Zugleich müssten überfüllte Curricula reduziert, Transfersituationen ausgewiesen und Curriculum, Unterrichtsentwicklung und Prüfung konsequent aufeinander abgestimmt werden.

7.3 Zusätzliche Kompetenznachweise erproben

Am Beispiel kanadischer *Credential*-Ansätze könnte geprüft werden, wie Abschlusszeugnisse besondere inner- und außerunterrichtliche Formen des Engagements sichtbar machen können. Dadurch würden Leistungen zertifiziert, die in klassischen Fachnoten kaum abgebildet werden, ohne sofort neue hochselektive Bewertungssituationen zu erzeugen.

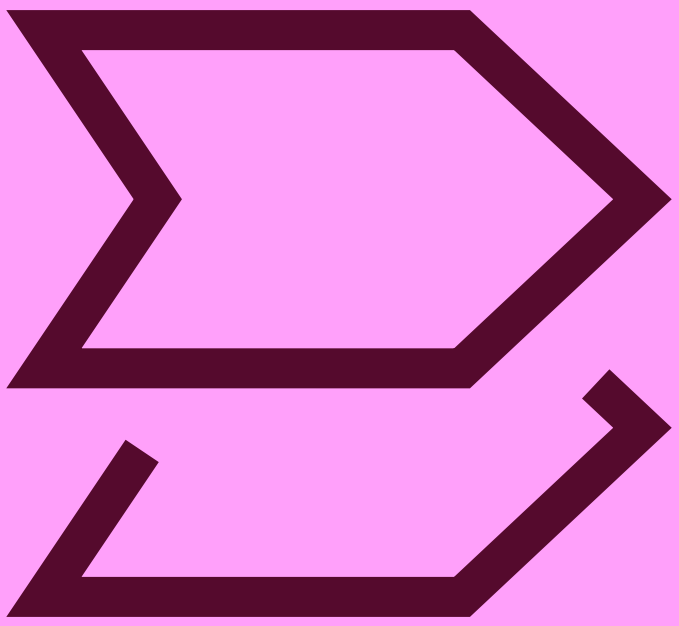
Eine mögliche Umsetzung bestünde in Zertifikaten für Projektarbeit, gesellschaftliches Engagement, Forschungsprojekte, Wettbewerbe oder besondere Beiträge zum schulischen Gemeinschaftsleben. Dazu könnten etwa die Mitarbeit in der Schüler:innenvertretung, die Organisation schulischer Veranstaltungen oder kulturelle Projekte zählen.

7.4 Zertifikatskurs zu *Assessment Literacy* entwickeln

Zuletzt gilt: Mehr Autonomie erfordert Handlungsfähigkeit. Im Sinne professioneller Autonomie, wie sie etwa in Finnland stark ausgeprägt ist, könnte ein bundesweit anschlussfähiger Zertifikatskurs dazu beitragen, Lehrkräfte, Schulleitungen und Schulaufsicht für die Entwicklung, Durchführung und Bewertung komplexer Prüfungsformate zu qualifizieren. Im Zentrum steht *Assessment Literacy*: also die Fähigkeit, Lernziele, Aufgabenformate, Bewertungsverfahren und Rückmeldungen kohärent aufeinander zu beziehen. *Legal Literacy* sollte darin als verbindlicher Bestandteil verankert werden, damit Lehrkräfte rechtssicher und gleichzeitig flexibel mit den bereits bestehenden Freiheiten umgehen können, um angemessene Prüfungen zu entwickeln.

> **Mehr Autonomie erfordert Handlungsfähigkeit.**

Eine mögliche Umsetzung bestünde in modularen Fortbildungen durch Hochschulen, Landesinstitute, Stiftungen und zivilgesellschaftliche Partner. Inhaltlich könnten Module zu Gütekriterien, Bewertungsrastern, Feedback, Kollaboration, KI-Nutzung und Prüfungsrecht entwickelt und als Fortbildung oder Zusatzqualifikation anerkannt werden.



8
Fazit

Im internationalen Vergleich zeigt sich, dass die Weiterentwicklung von Prüfungskulturen keine rein didaktische oder messtheoretische Frage ist, sondern eine zutiefst systemische und politische Herausforderung. Reformen greifen in etablierte Erwartungen an Vergleichbarkeit, Fairness und Selektionsfunktionen ein und stoßen deshalb nicht selten auf Widerstände von Eltern oder bildungspolitischen Akteur:innen, wie sich exemplarisch am Fall Singapurs beobachten lässt. Prüfungssysteme sind eng mit gesellschaftlichen Leistungsnormen und Zugangsmechanismen verknüpft und können daher nicht isoliert verändert werden.

Zugleich wird deutlich, dass auch international erfolgreiche Bildungssysteme keine vollständig pädagogisch und juristisch geklärte „Lösung" für die Bewertung komplexer Kompetenzen bereithalten. Sichtbar wird vielmehr eine gemeinsame Grundspannung: Je komplexer Kompetenzen konzipiert und erfasst werden, desto schwieriger ist es, die entsprechenden Messverfahren zu standardisieren, zu objektivieren und zu vergleichen. Damit geraten sie in ein Spannungsverhältnis zu den Anforderungen rechtssicherer, selektionswirksamer Abschlussprüfungen, die auf Transparenz, Vergleichbarkeit und Nachvollziehbarkeit angewiesen sind.

Vor diesem Hintergrund ist international eine doppelte Bewegung erkennbar. Einerseits entstehen Versuche, komplexe Kompetenzen auch in standardisierten Settings kognitiv zu operationalisieren, etwa im Rahmen von *Large-Scale Assessments* wie PISA. Diese Ansätze liefern wichtige Impulse für Modellierung und empirische Erfassung, stoßen jedoch an Grenzen, sobald sie als Grundlage hochschulzugangsrelevanter Prüfungsentscheidungen dienen sollen. Bislang existieren keine Verfahren, die den hierfür erforderlichen Ansprüchen an Validität, Reliabilität und rechtlicher Absicherung gleichermaßen gerecht werden.

> **Es zeigt sich in mehreren besonders hoch entwickelten Schulsystemen die Tendenz, Räume jenseits selektionswirksamer Prüfungen zu schaffen, in denen komplexe Kompetenzen vor allem im Sinne eines *Assessment for Learning* entwickelt und sichtbar gemacht werden.**

Andererseits zeigt sich in mehreren besonders hoch entwickelten Schulsystemen die Tendenz, Räume jenseits selektionswirksamer Prüfungen zu schaffen, in denen komplexe Kompetenzen vor allem im Sinne eines *Assessment for Learning* entwickelt und sichtbar gemacht werden. Formative Verfahren, projektbasierte Aufgabenformate und portfolioorientierte Ansätze ermöglichen es, Lernprozesse differenziert zu begleiten und individuelle Entwicklungsverläufe stärker in den Blick zu nehmen. Damit verschiebt sich der Fokus von einer wettbewerbsorientierten sozialen Bezugsnorm hin zu einer stärker individuellen, entwicklungsorientierten Perspektive auf Leistung. Dies zeigt sich in einem Bedeutungszuwachs

der Kombination aus kriterialer Bezugsnorm (*Learning Progressions*) und individueller Bezugsnorm, die die Lernenden auch bezüglich Motivation und Emotion beim Lernen in den Blick nimmt und unterstützt. Wie Brandt und Kolleg:innen betonen, handelt es sich hierbei um „formative assessment practices, which focus on developmental feedback without high-stakes consequences“ (Brandt et al. 2025, S. 4).

Zentral für die Umsetzung solcher Ansätze ist die professionelle Handlungskompetenz von Lehrkräften. Internationale Beispiele verdeutlichen, dass weniger die flächendeckende Integration komplexer Kompetenzmessungen in *High-Stakes*-Prüfungen im Vordergrund steht als vielmehr der Aufbau von *Assessment Literacy* und die Stärkung professioneller Autonomie. Daraus wurden für Deutschland vier mögliche Stellschrauben formuliert: den rechtlichen Rahmen für formative Leistungsnachweise erweitern, selektionsärmere Entwicklungsräume vor der Qualifikationsphase stärken, curriculare Freiräume im Sinne des *Constructive Alignment* schaffen sowie *Assessment Literacy* durch systematische Qualifizierung ausbauen.

Insgesamt legt der Blick auf besonders hoch entwickelte Schulsysteme nahe, dass eine nachhaltige Weiterentwicklung von Prüfungskulturen nicht primär in einer immer stärkeren Operationalisierung komplexer Kompetenzen für standardisierte Abschlussprüfungen liegt. Vielmehr zeigt sich ein Bewusstsein dafür, dass *High-Stakes*-Prüfungen Lernprozesse tendenziell verengen. Als Reaktion werden selektionswirksame Prüfungssituationen begrenzt und Freiräume für die Entwicklung komplexer Kompetenzen geschaffen. Die Herausforderung wird in den nächsten Jahren darin bestehen, die Selektionsfunktion der Leistungsbewertung und ihre Funktion in der Förderung komplexer Persönlichkeitsentwicklung in ein produktives Verhältnis zu setzen, ohne die jeweils spezifischen Anforderungen aus dem Blick zu verlieren.

Literatur

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A. & Persson, T. (2015). Strategies for teaching students to think critically: a meta-analysis. *Review of Educational Research, 85*(2), 275–314. https://doi.org/10.3102/0034654314551063.

Alberta Ministry of Education (2026a). *Graduation Requirements, Credentials and Credits*. https://www.alberta.ca/graduation-requirements-credentials-and-credits.

Alberta Ministry of Education (2026b). *High School Completion*. https://www.alberta.ca/high-school-completion.

Anderson, C. M., Summers, K. H., Kopatich, R. D. & Dwyer, W. B. (2023). Collective teacher efficacy and its enabling conditions: a proposed framework for influencing collective efficacy in schools. *AERA Open, 9*. https://doi.org/10.1177/23328584231175060.

Arnold, R. (2018). *Ich lerne, also bin ich: Eine systemisch-konstruktivistische Didaktik* (3. Aufl.). Carl-Auer.

Australian Curriculum, Assessment and Reporting Authority (2022). *F–10 Curriculum Overview*. https://www.australiancurriculum.edu.au/f-10-curriculum/f-10-curriculum-overview.

Australian Curriculum, Assessment and Reporting Authority (2026). *General Capabilities (Version 8.4)*. https://acara.edu.au/curriculum/foundation-year-10/general-capabilities.

Badrinarayan, A. (2024). *Instructionally Relevant Assessment Systems: What is the Role of Performance Assessments?* Learning Policy Institute. https://learningpolicyinstitute.org/product/instructionally-relevant-assessment-systems-report.

Badrinarayan, A. (2025). *Design Principles for Instructionally Relevant Assessment Systems*. Learning Policy Institute. https://learningpolicyinstitute.org/product/design-principles-assessment-brief.

Barta, A., Póka, T. & Tamas, B. (2024). The comparison of several factorial structures of the Cornell Critical Thinking Test Level Z. *Studia Universitatis Babeș-Bolyai Psychologia-Paedagogia, 69*(1), 5–26. https://doi.org/10.24193/subbpsyped.2024.1.01.

Barthakur, A., Viberg, O., Kizilcec, R. F., Baker, R. S. & Dawson, S. (2026). Advancing 21st-century professional competencies with learning analytics in the age of generative AI. *Journal of Learning Analytics, 13*(1), 1–6. https://doi.org/10.18608/jla.2026.9743.

Beutel, S.-I. (2024). *Lernen ohne Noten: Alternative Konzepte der Leistungsbeurteilung* (2. Aufl.). Kohlhammer.

Bialik, M. & Fadel, C. (2017). *Overcoming System Inertia in Education Reform*. Center For Curriculum Redesign. https://curriculumredesign.org/wp-content/uploads/Overcoming-Inertia-in-Education-Reform.pdf.

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education, 1*. https://www.tru.ca/__shared/assets/Constructive_Alignment36087.pdf.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M. & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw & E. Care (Hrsg.), *Assessment and Teaching of 21st Century Skills* (S. 17–66). Springer.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102.

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgment*. Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.110782.

Brandt, C., Evans, C. & Domaleski, C. (2025). *Assessing 21st Century Competencies: Guiding Principles for States and Districts.* Center for Assessment. https://www.nciea.org/wp-content/uploads/2024/12/Assessing21stCenturyCompetencies-Report-FINAL.pdf.

Breidenstein, G. (2018). Das Theorem der „Selektionsfunktion der Schule" und die Praxis der Leistungsbewertung. In S. Reh & N. Ricken (Hrsg.), *Leistung als Paradigma: Zur Entstehung und Transformation eines pädagogischen Konzepts* (S. 307–327). Springer Fachmedien. https://doi.org/10.1007/978-3-658-15799-9_14.

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D.-G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B. & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation, 5*(3). https://doi.org/10.59863/MIQL7785.

Bundesverfassungsgericht (2017). *Urteil vom 19. Dezember 2017* [Studienplatzvergabe im Fach Humanmedizin]. 1 BvL 3/14, 1 BvL 4/14. https://www.bundesverfassungsgericht.de/e/ls20171219_1bvl000314.html.

Care, E., Kim, H., Vista, A. & Anderson, K. (2018). *Education System Alignment for 21st Century Skills: Focus on Assessment*. Center for Universal Education at the Brookings Institution. https://www.brookings.edu/articles/education-system-alignment-for-21st-century-skills/.

College Board (2025). *AP Capstone Implementation Guide 2025–26*. https://apcentral.collegeboard.org/media/pdf/ap-capstone-implementation-guide.pdf.

Conley, D. T. & Darling-Hammond, L. (2013). *Creating Systems of Assessment for Deeper Learning*. Stanford Center for Opportunity Policy in Education.

Curriculum.nu (2018). *Handreiking brede vaardigheden*. https://curriculum.nu/download/Handreiking-brede-vaardigheden.pdf.

Darling-Hammond, L. & Adamson, F. (2010). *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*. Stanford Center for Opportunity Policy in Education (SCOPE). https://kammsolutions.com/wp-content/uploads/2015/05/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning_3.pdf.

Donohoo, J., O'Leary, T. & Hattie, J. (2020). The design and validation of the enabling conditions for collective teacher efficacy scale (EC-CTES). *Journal of Professional Capital and Community, 5*(2), 147–166. https://doi.org/10.1108/JPCC-08-2019-0020.

Erpenbeck, J., Rosenstiel, L., Grote, S. & Sauter, W. (2017). *Handbuch Kompetenzmessung: Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis*. Schäffer-Poeschel. https://doi.org/10.34156/9783791035123.

Eurydice Finland (2025). *Finland: Assessment in Single-Structure Education*. Eurydice. https://eurydice.eacea.ec.europa.eu/eurypedia/finland/assessment-single-structure-education.

Evans, C. M. (2019). Effects of New Hampshire's innovative assessment and accountability system on student achievement outcomes after three years. *Education Policy Analysis Archives, 27*(10), 1–34. https://eric.ed.gov/?id=EJ1204567.

Fadel, C., Black, A., Taylor, R., Slesinski, J. & Dunn, K. (2024). *Bildung für das Zeitalter von KI*. ZLL21.

Finland Ministry of Education and Culture (2026). *Basic Education 2045: For Life – A Vision for Finnish Comprehensive Schools*. https://urn.fi/URN:ISBN:978-952-415-061-3.

Finnish Education Evaluation Centre (2026). *Finnish Education Evaluation Centre* [FINNEC bzw. Karvi]. https://www.karvi.fi/en.

Finnish National Agency for Education (2026a). *Transversal Competences in Finnish General Upper Education*. https://www.oph.fi/en/education-and-qualifications/transversal-competences-finnish-general-upper-secondary-education.

Finnish National Agency for Education (2026b). *What Is General Upper Secondary Education?* https://www.oph.fi/en/education-and-qualifications/what-general-upper-secondary-education.

Finnish National Agency for Education (2026c). *Finnish Education in a Nutshell*. https://www.oph.fi/sites/default/files/documents/Finnish_education_in_a_nutshell_1.pdf.

Foster, N. & Piacentini, M. (Hrsg.) (2023). *Innovating Assessments to Measure and Support Complex Skills*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/04/innovating-assessments-to-measure-and-support-complex-skills_b0255009/e5f3e341-en.pdf.

Gan, S. H. (2025). *MOE FY2025 Committee of Supply Debate: Response by Minister of State for Education Gan Siow Huang*. Singapore Ministry of Education (6. März 2025). https://www.moe.gov.sg/news/speeches/20250306-moe-fy2025-committee-of-supply-debate-response-by-minister-of-state-for-education-gan-siow-huang.

Godau, M., Barreiro, J., Hermann, K., Neuhausen, T., Maxelon, D. & Weidner, V. (2025). „Wofür bin ich denn hier, wenn ich nicht in irgendeiner Weise auch helfen kann?" Zur Rolle der Lehrkraft beim selbstständigen Songwriting im Unterricht. In U. Frey-Konrad, V. C. Hofbauer, O. Blanchard & L. Bugiel (Hrsg.), *46. Jahresband des Arbeitskreises Musikpädagogische Forschung* (S. 167–180). Waxmann. https://doi.org/10.31244/9783818850715.13.

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C. & Germany, M.-L. (2018). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin & M. Wilson (Hrsg.), *Assessment and Teaching of 21st Century Skills: Research and Applications* (S. 75–91). Springer International Publishing. https://doi.org/10.1007/978-3-319-65368-6_5.

Griffin, P. & Care, E. (Hrsg.) (2015). *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Springer Dordrecht. https://doi.org/10.1007/978-94-017-9395-7.

Hao, J., Von Davier, A. A., Yaneva, V., Lottridge, S., Von Davier, M. & Harris, D. J. (2024). Transforming assessment: the impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice, 43*(2), 16–29. https://doi.org/10.1111/emip.12602.

Hardy, I., Hamid, M. O. & Reyes, V. (2021). The pragmatic perfectionism of educational policy: reflections on/from Singapore. *Globalisation, Societies and Education, 19*(3), 355–370. https://doi.org/10.1080/14767724.2020.1831905.

Hart, M. (2019). *Measuring What Matters: Testing a New Approach to Accountability*. People for Education.

Harvard University (o. J.). *Explore SEL* [Social and Emotional Learning]. http://exploresel.gse.harvard.edu.

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge. https://doi.org/10.4324/9780203887332.

Hattie, J. (2023). *Visible Learning, the Sequel: A Synthesis of over 2,100 Meta-Analyses Relating to Achievement*. Routledge. https://doi.org/10.4324/9781003380542.

Hautamäki, J. & Kupiainen, S. (2014). Learning to learn in Finland. In R. Deakin Crick, C. Stringher & K. Ren (Hrsg.), *Learning to Learn: International Perspectives from Theory and Practice* (25 S.). Routledge. https://doi.org/10.4324/9780203078044-9.

Helmke, A. (2022). *Unterrichtsqualität und Professionalisierung: Diagnostik von Lehr-Lern-Prozessen und evidenzbasierte Unterrichtsentwicklung* (umf. akt. Neuausgabe). Klett Kallmeyer.

Hesse, F., Care, E., Buder, J., Sassenberg, K. & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Hrsg.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (S. 37–56). Springer Dordrecht.

Hipkins, R. & Cameron, M. (2018). *Trends in Assessment: An Overview of Themes in the Literature*. New Zealand Council for Educational Research. https://www.nzcer.org.nz/sites/default/files/downloads/Trends%20in%20assessment%20report.pdf.

Hurrelmann, K. (2020). Nicht mehr zeitgemäß fürs 21. Jahrhundert: Ändert das Abitur – und zwar radikal! *Tagesspiegel* (2. Juli 2020). https://www.tagesspiegel.de/wissen/andert-das-abitur--und-zwar-radikal-4179163.html.

International Baccalaureate (2023). *UWC Atlantic and International Baccalaureate Partner to Transform Education*. https://www.ibo.org/news/news-about-ib-schools/uwc-atlantic-and-international-baccalaureate-partner-to-transform-education/.

International Baccalaureate (2024). *IB Launches New Pilot Course with UWCSEA for August 2024*. https://www.ibo.org/news/news-about-the-ib/ib-launches-new-pilot-course-with-uwcsea-for-august-2024/.

International Baccalaureate (2025). *Artificial Intelligence in IB Assessment and Education: A Crisis or an Opportunity?* https://ibo.org/news/news-about-the-ib/artificial-intelligence-in-ib-assessment-and-education-a-crisis-or-an-opportunity2/.

International Baccalaureate (2026a). *Academic Integrity*. https://ibo.org/about-the-ib/what-it-means-to-be-an-ib-student/responsibilities-of-students-and-ib-world-schools/academic-integrity/.

International Baccalaureate (2026b). *Systems Transformation*. https://ibo.org/programmes/collaborative-review-of-the-dp-and-cp/systems-transformation/.

International Baccalaureate (2026c). *Collaborative Review of the DP and CP (16+ Review)*. https://www.ibo.org/programmes/collaborative-review-of-the-dp-and-cp/.

Joarder, S. & Chatti, M. A. (2026). Human-centred development of indicators for self-service learning analytics: a transparency through exploration approach. *Journal of Learning Analytics, 13*(1), 163–189. https://doi.org/10.18608/jla.2026.8921.

Jürgens, E. & Lissmann, U. (2015). *Pädagogische Diagnostik: Grundlagen und Methoden der Leistungsbeurteilung in der Schule*. Beltz.

Kain, C., Koschmieder, C. & Matischek-Jauk, M. (2024). Mapping the landscape: a scoping review of 21st century skills literature in secondary education. *Teaching and Teacher Education, 151*, Artikel 104739. https://www.sciencedirect.com/science/article/pii/S0742051X24002713.

Koh, E., Jonathan, C., Lim, F. V., Kadir, M. & Tan, J. P.-L. (2023). Value and challenges in using a collaborative critical reading and learning analytics system: a cross-case analysis of two high schools. In C. Damșa, M., Borge, E. Koh & M. Worsley (Hrsg.), *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning – CSCL 2023* (S. 233–236). International Society of the Learning Sciences. https://doi.org/10.22318/cscl2023.869864.

Kultusministerkonferenz (2024). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung (Beschluss der KMK vom 07.07.1972 i. d. F. vom 06.06.2024)*. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1972/1972_07_07-VB-gymnasiale-Oberstufe-Abiturpruefung.pdf.

Lam, K., Shin, L. W., Pang, A. & Cheang, T. K. (2025). Assessment and examination policies in the Singapore education system. In K. L. Chue, A. Yeo, K. H. K. Tan & T. H. Yong (Hrsg.), *Fundamentals of Assessment: Principles and Practices for the Classroom* (S. 17–30). Routledge. https://doi.org/10.4324/9781003534648-3.

Leach, S. M., Immekus, J. C., French, B. F. & Hand, B. (2020). The factorial validity of the Cornell Critical Thinking Tests: a multi-analytic approach. *Thinking Skills and Creativity, 37*, Artikel 100676. https://doi.org/10.1016/j.tsc.2020.100676.

López-Crespo, G., Blanco-Gandía, M. C., Valdivia-Salas, S., Fidalgo, C. & Sánchez-Pérez, N. (2022). The educational e-portfolio: preliminary evidence of its relationship with student's self-efficacy and engagement. *Education and Information Technologies, 27*(4), 5233–5248. https://doi.org/10.1007/s10639-021-10827-2.

Lorié, W. & Dadey, N. (2026). *Artificial Intelligence in Large-Scale Assessment Programs: Applications and Considerations for State Education Agencies*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/library/artificial-intelligence-in-large-scale-assessment-systems/.

McGrane, J. A., Johnston, S.-K., Vendrell i Morancho, M. & Hopfenbeck, T. N. (2022). *Critical Thinking Development in the IB Middle Years Programme: An International Multisite Evaluation*. University of Oxford Department of Education. https://ibo.org/globalassets/new-structure/research/pdfs/myp-critical-thinking-final-report-en.pdf.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. https://doi.org/10.1177/026553229601300302.

Meyer, H.-D. & Zahedi, K. (2014). *Nein zu Pisa (Offener Brief an Andreas Schleicher)*. Gesellschaft für Bildung und Wissen. https://bildung-wissen.eu/fachbeitraege/nein-zu-pisa-offener-brief-an-andreas-schleicher.html.

Mislevy, R. J., Russell, G. A. & Janice, F. L. (2004). *A Brief Introduction to Evidence-Centered Design. CSE Report 632*. National Center for Research on Evaluation. https://files.eric.ed.gov/fulltext/ED483399.pdf.

MBL Work Group (2021). *Mastery-Based Learning in Washington State. 2021 Report by the Mastery-based Learning Work Group*. Washington State Board of Education. https://sbe.wa.gov/sites/default/files/2024-08/2021%20MBL%20Work%20Group%20Report.pdf.

New Zealand Ministry of Education (2007). *The New Zealand Curriculum: For English-Medium Teaching and Learning in Years 1–13*. Learning Media.

New Zealand Qualifications Authority (o. J.). *About Standards*. https://www2.nzqa.govt.nz/ncea/about-ncea/about-standards/.

NextEducation (o. J.). *Future Skills im Vergleich: Kompetenzen für die Zukunft verstehen, einordnen und gestalten*. https://futureskills4u.org/de/.

Niedersächsisches Kultusministerium (2015). *Die Arbeit in den Schuljahrgängen 5 bis 10 des Gymnasiums. RdErl. d. MK v. 23.6.2015 – 33-81011 – VORIS 22410*. https://bildungsportal-niedersachsen.de/index.php?eID=dumpFile&t=f&f=11560&token=d3e9a53e0c87ceef8225495989b612909e4fa219.

OECD (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/11/pisa-2015-results-volume-v_g1g83e07/9789264285521-en.pdf.

OECD (2020). *OECD Lernkompass 2030. OECD-Projekt Future of Education and Skills 2030. Rahmenkonzept des Lernens*. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/OECD_Lernkompass_2030.pdf.

OECD / European Commission (2026). *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*. OECD Publishing. https://ailiteracyframework.org/pdfs/framework_pdf/AILF_en.pdf.

Ontario Ministry of Education (2010). *Collaborative Teacher Inquiry: New Directions in Professional Practice* (Capacity Building Series, Secretariat Special Edition Nr. 16). https://www.onted.ca/monographs/capacity-building-series/collaborative-teacher-inquiry.

Ontario Ministry of Education (2020). *Earning Your High School Diploma*. https://www.ontario.ca/page/earning-your-high-school-diploma.

Ontario Ministry of Education (2022a). *Online Learning for Secondary Students*. https://www.ontario.ca/page/online-learning-secondary-students.

Ontario Ministry of Education (2022b). *Policy/Program Memorandum 167*. https://www.ontario.ca/document/education-ontario-policy-and-program-direction/policyprogram-memorandum-167.

Ontario Ministry of Education (2024a). *Get Your High School Volunteer Hours*. https://www.ontario.ca/page/get-your-high-school-volunteer-hours.

Ontario Ministry of Education (2024b). *Transferable Skills*. https://www.dcp.edu.gov.on.ca/en/program-planning/transferable-skills/introduction.

Ouyang, F., Xu, W. & Cukurova, M. (2023). An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective. *International Journal of Computer-Supported Collaborative Learning, 18*(1), 39–66. https://doi.org/10.1007/s11412-023-09387-z.

Pellegrino, J. W. & Hilton, M. L. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. National Academies Press.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354.

Rampelt, F., Matthes, W., Hannken-Illjes, K., Sandmeir, A., Gehrs, V., Horstmann, N., Kunz, A. M., Eigbrecht, L., Johannsen, T., Blum, S., Frank, S., Sutter, C. & Koch, H. (2025). *Future Skills 2030. Ein aktualisiertes Framework für Zukunftskompetenzen*. Zenodo. https://doi.org/10.5281/ZENODO.17910748.

Rampelt, F., Matthes, W., Röwert, R., Buck, I., Winde, M. & Koch, H. (2024). *Future-Skills-Charta: Wir stärken Kompetenzen für eine Welt im Wandel*. Stifterverband. https://doi.org/10.5281/ZENODO.14499573.

Reh, S., Bühler, P., Hofmann, M. & Moser, V. (Hrsg.) (2021). *Schülerauslese, schulische Beurteilung und Schülertests 1880–1980*. Julius Klinkhardt. https://doi.org/10.35468/5890.

Rheinberg, F. & Engeser, S. (2018). Intrinsische Motivation und Flow-Erleben. In J. Heckhausen & H. Heckhausen, *Motivation und Handeln* (5. Aufl., S. 423–450). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-53927-9_14.

Runge, A., Attali, Y., LaFlair, G. T., Park, Y. & Church, J. (2024). A generative AI-driven interactive listening assessment task. *Frontiers in Artificial Intelligence, 7*, Artikel 1474019. https://doi.org/10.3389/frai.2024.1474019.

Rutkowski, L. & Rutkowski, D. (2025). Methodological reflections on PISA's creative thinking assessment. *The Journal of Creative Behavior, 59*(2), Artikel e70038. https://doi.org/10.1002/jocb.70038.

Schaffar, B. & Wolff, L.-A. (2024). Phenomenon-based learning in Finland: a critical overview of its historical and philosophical roots. *Cogent Education, 11*(1). https://doi.org/10.1080/2331186X.2024.2309733.

Shepard, L., Daro, P. & Stancavage, F. B. (2013). *The Relevance of Learning Progressions for NAEP*. American Institutes for Research. https://eric.ed.gov/?id=ED545240.

Shukur, I. (2025). Assessing global perspectives on the IB curriculum: a qualitative study of international teachers. *OTS Canadian Journal, 4*(5), 31–50. https://doi.org/10.58840/qr7w1594.

Singapore Ministry of Education (2024a). *21st Century Competencies*. https://www.moe.gov.sg/education-in-sg/21st-century-competencies.

Singapore Ministry of Education (2024b). *Initiatives to Engage Parents to Alleviate Pressure of PSLE for Mental Well-Being of Students and Parents*. https://www.moe.gov.sg/news/parliamentary-replies/20241111-initiatives-to-engage-parents-to-alleviate-pressure-of-psle-for-mental-well-being-of-students-and-parents.

Singapore Ministry of Education (2025). *Digital Literacy and Technological Skills*. https://www.moe.gov.sg/education-in-sg/educational-technology-journey/edtech-masterplan/digital-literacy-and-technological-skills.

Singapore Ministry of Education (2026). *Artificial Intelligence in Education*. https://www.moe.gov.sg/education-in-sg/educational-technology-journey/edtech-masterplan/artificial-intelligence-in-education.

Singapore Student Learning Space (2025). *What's New in December 2025 Update* (Release 3). https://www.learning.moe.edu.sg/latest-news/whats-new-dec2025/.

Sliwka, A. & Klopsch, B. (2022). *Deeper Learning in der Schule: Pädagogik des digitalen Zeitalters*. Julius Beltz.

Suat Hoon, P., Yoke Peng, L., Koh, E. & Siew Kee, C. (2025). The scaling of a teamwork innovation in a Singapore tertiary institution: value orchestration and co-creation. *Asia Pacific Journal of Education*, 1–20. https://doi.org/10.1080/02188791.2025.2571641.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N. & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, Artikel 100075. https://doi.org/10.1016/j.caeai.2022.100075.

Tan, C. (2024). An analysis of attainment grouping policy in Singapore. *British Educational Research Journal, 51*(1), 394–415. https://doi.org/10.1002/berj.4080.

Tan, J. P.-L., Koh, E., Chan, M., Costes-Onishi, P. & Hung, D. (2017). *Advancing 21st Century Competencies in Singapore*. Asia Society Center for Global Education.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20–27. https://doi.org/10.1080/00405840802577569.

Undervisningsministeriet (2024). *Ekspertgruppen om ChatGPT og andre digitale hjælpemidler*. https://uvm.dk/media/autbhhja/240423-ekspertgruppen-om-chatgpt-og-andre-digitale-hjaelpemidlers-rapport.pdf.

Undervisningsministeriet (2026). *Ny strakspakke mod AI-snyd på gymnasierne*. https://uvm.dk/aktuelt/nyheder/2026/august/260806-ny-strakspakke-mod-ai-snyd-paa-gymnasierne/.

UNESCO (2023). *Guidance for Generative AI in Education and Research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research.

University of Melbourne (2025). *Competency Kit: Assessing Agency in Learning*. https://education.unimelb.edu.au/__data/assets/pdf_file/0008/5199092/2025-Competency-Kit_AGENCY-IN-LEARNING_generic_Sample.pdf.

University of Melbourne (o. J.). *Melbourne Metrics*. https://education.unimelb.edu.au/melbourne-metrics/about.

UWC South East Asia (o. J.). *IB Systems Transformation: Leadership for Change*. https://www.uwcsea.edu.sg/ib-systems-transformation.

Verheyen, N. (2018). *Die Erfindung der Leistung*. Hanser Berlin.

Vygotsky, L. S. (2021). *L. S. Vygotsky's Pedological Works. Volume 2: The Problem of Age* (Perspectives in Cultural-Historical Research, Band 10). Springer Singapore. https://doi.org/10.1007/978-981-16-1907-6.

Wampfler, P. (2024a). Eigenständigkeit, Vergleichbarkeit, Allokation. In A. Langela-Bickenbach, R. Dreier, P. Wampfler & C. Albrecht (Hrsg.), *Wege zu einer zeitgemäßen Prüfungskultur: Grundlagen und Praxisbeispiele* (S. 39–42). Julius Beltz.

Wampfler, P. (2024b). Meritokratie und individuelle Leistung als Märchen für Erwachsene. In A. Langela-Bickenbach, R. Dreier, P. Wampfler & C. Albrecht (Hrsg.), *Wege zu einer zeitgemäßen Prüfungskultur: Grundlagen und Praxisbeispiele* (S. 34–38). Julius Beltz.

Whyte, A. & Nael, M. (2025). Ministry of Education developing a new exam information system. *ERR News* (22. Januar 2025). https://news.err.ee/1609583167/ministry-of-education-developing-a-new-exam-information-system.

Wong, H. M., Kwek, D. & Tan, K. (2020). Changing assessments and the examination culture in Singapore: a review and analysis of Singapore's assessment policies. *Asia Pacific Journal of Education, 40*(4), 433–557. https://doi.org/10.1080/02188791.2020.1838886.

Woolley, A. W., Aggarwal, I. & Malone, T. W. (2021). Collective intelligence and group performance: introduction to the Special Issue. *Computers in Human Behavior, 119*, Artikel 106699.

Impressum

Herausgegeben von
Robert Bosch Stiftung GmbH

Autor:innen
Prof. Dr. Anne Sliwka
Dominik Maxelon
Jule Eckert
Institut für Bildungswissenschaft, Universität Heidelberg

Verantwortlich bei der Robert Bosch Stiftung
Carolin Genkinger, carolin.genkinger@bosch-stiftung.de
Vincent Steinl, vincent.steinl@bosch-stiftung.de

Lektorat
Julia Kühn, Berlin

Gestaltung
mischen, www.mischen-berlin.de

Designmanagement
Jonas Wilisch
Julia Wittich

Copyright 2026
Robert Bosch Stiftung GmbH, Stuttgart
Alle Rechte vorbehalten

Robert Bosch Stiftung GmbH
Heidehofstraße 31
70184 Stuttgart
www.bosch-stiftung.de

Zitiervorschlag
Sliwka, Anne; Maxelon, Dominik; Eckert, Jule (2026): *Prüfungskulturen im Wandel. Ein internationaler Überblick zu Forschungsstand, Policy und Praxis*. Stuttgart: Robert Bosch Stiftung

Über die Robert Bosch Stiftung
Die Robert Bosch Stiftung arbeitet in den Bereichen Gesundheit, Bildung und Globale Fragen. Mit ihrer Förderung setzt sie sich für eine gerechte und nachhaltige Zukunft ein. Die Stiftung ist gemeinnützig, unabhängig und überparteilich. Sie geht auf das Vermächtnis von Robert Bosch zurück. Der Unternehmer und Stifter formulierte darin den doppelten Auftrag, das Fortbestehen des Unternehmens zu sichern und sein soziales Engagement weiterzuführen. Die Robert Bosch Stiftung GmbH unterhält eigene Einrichtungen, entwickelt innovative Projekte und fördert auf internationaler wie lokaler Ebene. Die Erkenntnisse aus ihrer Förderung bringt die Stiftung in die Fachwelt und die öffentliche Debatte ein.