ERWEITERTE ZUSAMMENFASSUNG

INNOVATIVE ANSÄTZE ZUR MESSUNG UND FÖRDERUNG KOMPLEXER FÄHIGKEITEN

Ursprünglich auf Englisch veröffentlicht von

OECD

Bildnachweis: Umschlaggestaltung auf der Grundlage von Bildern von © Shutterstock/treety; © Shutterstock/Merfin..

ÜBER DIESE VERÖFFENTLICHUNG

Die vorliegende Broschüre fasst die Kernaussagen der OECD-Publikation *Innovating Assessments to Measure and Support Complex Skills* (Foster und Piacentini, 2023) zusammen. *Innovating Assessments* ist das Ergebnis einer Zusammenarbeit zwischen dem OECD-Sekretariat und der PISA-Forschungs- und Innovationsgruppe (RIG) sowie mehreren anderen internationalen Experten und Mitarbeitern auf dem Gebiet der Bildungsmessung und des Testdesigns.

Die Veröffentlichung der vorliegende Broschüre wurde durch die Unterstützung der Bertelsmann Stiftung, der Deutsche Telekom Stiftung, der Stiftung Mercator sowie der Robert Bosch Stiftung ermöglicht.

Natalie Foster und Mario Piacentini haben den ursprünglichen Band herausgegeben und mehrere Kapitel beigesteuert. RIG-Mitglieder, darunter Kadriye Ercikan, Xiangen Hu, Cesar A. Amaral Nunes, James Pellegrino, Ido Roll und Kathleen Scalise, sowie eingeladene Mitarbeiter, darunter Miri Barhak-Rabinowitz, Hongwen Guo, Han Hui Por, Errol Kaylor, Cassie Malcom, Argenta Price, John. P. Sabatini, Keith Shubeck und Carl Wieman trugen zu den übrigen Kapiteln bei und gaben fachkundigen Rat und Feedback zur gesamten Veröffentlichung. Andreas Schleicher, OECD-Direktor für Bildung und Qualifikationen, und Yuri Belfali, Leiter der Abteilung Frühkindliche Entwicklung und Schulen bei der OECD, gaben zusätzliche Hinweise und Rückmeldungen. Die vorliegende Broschüre wurde von Mario Piacentini, Natalie Foster und Marc Fuster (OECD) erstellt.

Foster, N. und M. Piacentini (Hrsg.) (2023), *Innovating Assessments to Measure and Support Complex Skills* - Extended executive summary, OECD Publishing, Paris, https://www.oecd.org/pisa/innovation.

INHALTSVERZEICHNIS

ÜBER DIESE VERÖFFENTLICHUNG 3

INHALTSVERZEICHNIS 4

EDITORIAL 6

LEISTUNGSMESSUNG IST WICHTIG 9

VERLAGERUNG DER BILDUNGSZIELE: DER SCHWERPUNKT LIEGT AUF DEN KOMPETENZEN DES 21. JAHRHUNDERTS 10

WORUM GEHT ES BEI DEN KOMPETENZEN DES 21. JAHRHUNDERTS? 11

DIE MESSUNG DER KOMPETENZEN DES 21. JAHRHUNDERTS ERFORDERT EIN INNOVATIVES KONZEPT 12

LEISTUNGSMESSUNGEN DER NÄCHSTEN GENERATION: DESIGNPRINZIPIEN UND BEISPIELE 14

LEISTUNGSMESSUNG ALS EIN PROZESS DER INFERENZ AUS BELEGEN 14

ERNEUERUNG DER KOGNITIONSKOMPONENTE: DEFINITION VON KONSTRUKTEN DER KOMPETENZMESSUNG 18

FRÜHZEITIGE ENTSCHEIDUNGEN ÜBER DEN SCHWERPUNKT DER MESSUNG DER KOMPETENZEN DES 21. JAHRHUNDERTS 18
Relevante Aktivitäten zur Bewertung der Kompetenzen des 21. Jahrhunderts 19
Praxiskontexte oder Anwendungsbereiche 22
Individuelle vs. kollaborative Aufgaben 24
SCHAFFUNG SOLIDER KONZEPTIONELLER GRUNDLAGEN 26
BERÜCKSICHTIGUNG SOZIOKULTURELLER UNTERSCHIEDE BEI DER DEFINITION VON BEWERTUNGSKONSTRUKTEN 29

ERNEUERUNG DER BEOBACHTUNGSKOMPONENTE: EINBEZUG VIELFÄLTIGERER UND INTERAKTIVER AUFGABEN 31

AUFGABENDESIGN NEU ÜBERDENKEN 31
Designprinzip I: Erweiterte, performanzbezogene „Low-Floor-High-Ceiling"-Aufgaben 32
Designprinzip II: Explizite Berücksichtigung von Fachwissen 37
NUTZUNG MODERNER TECHNOLOGIEN FÜR EIN INNOVATIVES DESIGN VON LEISTUNGSMESSUNGEN 38
Aufgabenformat: Von statischen zu interaktiven und dynamischen Bewertungssituationen 38
NEUE QUELLEN FÜR EVIDENZ: PRODUKT- UND PROZESSDATEN 45
Die Entstehung von Antwortprozessdaten 46

ERNEUERUNG DER INTERPRETATIONSKOMPONENTE: BEOBACHTUNGEN RICHTIG VERSTEHEN 48

EIN PRINZIPIENORIENTIERTER ANSATZ, UM KOMPLEXE DATEN ZU DEUTEN: EVIDENZREGELN UND STATISTIK BEI LEISTUNGSMESSUNGEN 48
Evidenzregeln 48
Auswahl eines geeigneten statistischen Modells 51
EINE GESCHICHTE AUS ZWEI WELTEN: ANSÄTZE DES MASCHINELLEN LERNENS UND EVIDENZBASIERTES DESIGN 51

DIE GRÜNDE FÜR KOMPLEXERE AUFGABEN UND PRAKTISCHE MÖGLICHKEITEN, SIE IN DER DOKUMENTATION ZU NUTZEN 55

INNOVATIVE LEISTUNGSMESSUNG: AUSBLICK 57

INVESTITIONEN IN LEISTUNGSMESSUNGEN DER NÄCHSTEN GENERATION 57

INTELLEKTUELLES KAPITAL 57
FINANZIELLES KAPITAL 59
POLITISCHES KAPITAL 60

INTERNATIONALE GROSS ANGELEGTE LEISTUNGSMESSUNGEN: MÖGLICHKEITEN FÜR INNOVATION IN GROSSEM MASSSTAB 61

PISA 2025 – LERNEN IN DER DIGITALEN WELT 61

CODA: ZURÜCK ZU DEN DREI ARTEN VON KAPITAL 64

REFERENZEN 65

EDITORIAL

Mehr als 20 Jahre nach ihrer Erstauflage hat sich die PISA-Studie (*Programme for International Student Assessment*) als eine entscheidende Triebkraft für Reformen im Bildungsbereich etabliert. Der bahnbrechende Gedanke der Studie bestand darin, die Fähigkeiten von Schüler/innen unmittelbar auf der Grundlage internationaler Maßstäbe zu testen, dies mit Daten von Schüler/innen, Lehrpersonen, Schulen und Systemen zu verknüpfen, um Leistungsunterschiede zu verstehen und die Potenziale internationaler Zusammenarbeit auszuschöpfen, um auf dieser Datenbasis Handlungsentscheidungen zu treffen.

PISA unterschied sich von Anfang an von traditionellen Formen der Leistungsmessung. Um bei PISA gut abzuschneiden, mussten die Schüler/innen in der Lage sein, von ihrem aktuellen Wissensstand zu extrapolieren, in ihrem Denken Fächergrenzen zu überschreiten und ihr Wissen kreativ in neuartigen Situationen anzuwenden, anstatt hauptsächlich das im Unterricht erworbene Wissen zu reproduzieren. Die moderne Welt belohnt uns nicht mehr für das, was wir wissen, sondern für das, was wir mit dem, was wir wissen, tun können. Da Inhalte immer leichter zugänglich werden und immer mehr kognitive Routineaufgaben digitalisiert und ausgelagert werden, muss sich der Schwerpunkt darauf verlagern, Menschen zum lebenslangen Lernen zu befähigen. Epistemisches Wissen – ein Verständnis davon, wie Wissenschaftler oder Mathematiker denken – und entsprechende Arbeitsweisen haben zunehmend Vorrang gegenüber der Kenntnis bestimmter Formeln, Namen oder Orte.

Diese Vision von Bildung spiegelt sich in vielen zeitgenössischen Rahmenwerken wider, welche die Ausbildung so genannter *21st century skills* fordern, einschließlich des Lernkompasses 2030 der OECD. Doch ohne grundlegende Veränderungen in unseren Bildungssystemen wird sich die Kluft zwischen dem, was diese unseren jungen Menschen bieten, und dem, was unsere Gesellschaft erfordert, wahrscheinlich noch weiter vergrößern.

Ein wesentlicher Bestandteil von Bildungssystemen sind Leistungsmessung und -kontrolle. Die Art und Weise, wie Schülerleistungen gemessen werden, hat aufgrund ihrer Signalwirkung für curriculare und unterrichtliche Schwerpunktsetzungen einen großen Einfluss auf die Zukunft der Bildung. Tests werden immer unser Denken darüber bestimmen, was wichtig ist, und das sollten sie auch – Lehrpersonen und Schulverwaltungen sowie Schüler/innen werden ihre Aufmerksamkeit auf die Inhalte von Tests richten und sich entsprechend anpassen. Eine grundlegende Frage ist, wie wir die Leistungsmessung angemessen gestalten und sicherstellen können, dass sie Lehrpersonen und politischen Entscheidungsträgern dabei hilft, den Fortschritt im Bildungswesen auf sinnvolle Weise zu messen.

Problematisch ist, dass viele Systeme der Leistungsmessung schlecht auf das Curriculum und auf die Kenntnisse und Fähigkeiten abgestimmt sind, die junge Menschen benötigen, um erfolgreich zu sein.

Beim Entwerfen von Assessments tauschen wir oft Gewinne an Gültigkeit und Relevanz gegen Gewinne an Effizienz und Zuverlässigkeit aus.

Solche Schwerpunktsetzungen haben jedoch ihren Preis: Der zuverlässigste und effizienteste Test ist ein solcher, bei dem die Schüler/innen auf eine Art und Weise antworten, die wenig Spielraum für Mehrdeutigkeit zulässt – typischerweise ein Multiple-Choice-Format. Ein relevanter Test ist ein solcher, bei dem wir ein breites Spektrum an Wissen und Fähigkeiten abfragen, die für den Erfolg in Leben und Beruf als wichtig erachtet werden.

Eine gute Leistungsmessung erfordert mehrere Antwortformate, einschließlich offener Formate, die komplexere Antworten ermöglichen. Solche Verfahren verlangen zwangsläufig komplexere Aus- und Bewertungsverfahren. Gute Tests sollten auch einen Einblick in die Denkweise und das Verständnis der Schüler/innen geben, indem sie deren Problemlösungsstrategien aufdecken und ihnen zugleich ein Feedback in angemessener Detailtiefe liefern, um ihre Entscheidungen zur Verbesserung zu unterstützen. Digitale Tests, die nicht nur die Antworten, sondern auch die Handlungen der Schüler/innen aufzeichnen, bieten verschiedene Möglichkeiten, die Leistungsmessung in diesem Sinne zu verbessern.

Darüber hinaus müssen die Formen der Leistungsmessung fair sein und eine adäquate Messung auf verschiedenen Detailebenen gewährleisten, damit sie für Entscheidungen auf verschiedenen Ebenen des Bildungssystems genutzt werden können. Wir müssen auch stärker daran arbeiten, die Kluft zwischen summativen und formativen Leistungsüberprüfungen zu überbrücken. Formale Bildung hat ihre Ursprünge in der handwerklichen Lehre, in der die Lehrlinge von und mit anderen Menschen lernten und eine unmittelbare und persönliche Rückmeldung über ihre Fortschritte erhielten. Jahrhunderte später hat die Industrialisierung des Bildungswesens das Lernen dann von der Leistungsmessung abgekoppelt, indem sie von Schülern verlangte, jahrelang zu lernen, um dann viel später dazu aufzurufen, das Gelernte in einem oft engen und zeitlich begrenzten Rahmen zu reproduzieren. Dies hat zu einem Lernen und Lehren beigetragen, das oft oberflächlich ist und sich auf das konzentriert, was sich leicht messen lässt. Die Digitalisierung bietet uns nun die Möglichkeit, summative und formative Elemente der Leistungsmessung zu kombinieren und kohärente, mehrschichtige Systeme der Leistungsmessung zu schaffen, die sich von den Schüler/innen selbst über die Klassenzimmer und Schulen bis hin zur regionalen, nationalen und sogar internationalen Ebene erstrecken. Eine bessere Integration von Bewertung und Lernen wird dazu führen, dass Lehrkräfte Tests nicht mehr als einen Zeitverlust beim Lernen betrachten, sondern als ein Instrument, das zum Lernen beiträgt.

All dies gilt natürlich auch für PISA. PISA wird weltweit als wichtiger Maßstab für den Erfolg von Schulsystemen wahrgenommen und muss als solcher die Bildungsreform anführen. Dank der Einführung der computergestützten Durchführung hat PISA seit 2012 sein Spektrum an Messgrößen erweitert und in jedem Zyklus einen neuen interdisziplinären Bereich aufgenommen – darunter Problemlösen (2012), kollaboratives Problemlösen (2015), globale Kompetenz (2018) und seit kurzem auch kreatives Denken (2022).

Im Jahr 2020 ging PISA noch einen Schritt weiter: Trotz der äußerst

schwierigen globalen Umstände beschlossen die Länder, mehr Ressourcen in die Entwicklung innovativer Formen der Leistungsmessung zu investieren, und richteten ein neues Programm für Forschung, Entwicklung und Innovation (FEI) ein, das von einer Gruppe hochrangiger internationaler Expert/innen für Leistungsmessung geleitet wird.

Die vorliegende Veröffentlichung ist das Ergebnis unserer Zusammenarbeit mit verschiedenen Expert/innen in den letzten drei Jahren seit Beginn unseres laufenden Forschungsprogramms. Sie zeigt auf, warum wir die Leistungsmessung innovativ gestalten müssen, erklärt, was wir ändern müssen und wie wir die Technologie nutzen können, um dieses Ziel zu erreichen. Er macht auch deutlich, dass dieser Wandel nicht über Nacht geschehen wird: Es gibt noch viel zu tun, und für eine großflächige Umsetzung dieser Ideen ist eine Bündelung politischen, finanziellen und intellektuellen Kapitals erforderlich.

PISA kann zu einem Motor werden, der diesen Wandel vorantreibt, indem die Potenziale internationaler Zusammenarbeit zwischen Pädagog/innen, Forschenden und politischen Entscheidungsträgern genutzt und sowohl die finanziellen als auch die politischen Kosten bei der Suche nach innovativen Praktiken zwischen den beteiligten Ländern aufgeteilt werden. Forschung und Innovation im Bereich umfassender Leistungsmessung sind seit jeher ein Kernbestandteil der PISA-DNA, und wir sind entschlossen, auf dem vor uns liegenden Weg weiterhin eine weltweit führende Rolle zu spielen.

Andreas Schleicher
Direktor des Direktorats für Bildung
Sonderberater für Bildungspolitik des Generalsekretärs

WAS SPRICHT FÜR *INNOVATIVE LEISTUNGSMESSUNG?*

Die vorliegende Broschüre fasst die Kernaussagen der OECD-Publikation *Innovating Assessments* (Foster und Piacentini, 2023) zusammen, die das Ergebnis einer mehrjährigen Forschungsarbeit internationaler Expert/innen auf dem Gebiet der Leistungsmessung und -bewertung und des OECD-Sekretariats ist.

Der Ausgangspunkt für diese Arbeit – und damit auch die argumentative Untermauerung von innovativer Leistungsmessung – sind eine Reihe miteinander verbundener Thesen. Zunächst einmal sollte uns die Leistungsmessung ein zentrales Anliegen sein. Sie ist ein wichtiger Wegweiser, der Schüler/innen zeigt, was sie lernen sollten und was sie können. Als solche sind sie eng mit Curricula und Pädagogik verknüpft und können Veränderungen in Bildungszielen und -praktiken vorantreiben oder verzögern. Die zweite These ergibt sich aus der ersten: Die Bildungsbewertung sollte sich auf das konzentrieren*, worauf es ankommt.* Die Frage, was wissenswert ist, was zu tun ist oder wie das Sein gestaltet werden kann, ist Gegenstand ständiger Debatten, wobei ein globales Narrativ dazu aufruft, das, was in der Schule gelehrt und gelernt wird, zu überdenken, um die Schüler/innen besser auf ihre Rolle als Bürger/innen und künftige Berufstätige vorzubereiten. Was diese beiden Thesen miteinander verbindet, ist der Gedanke, dass jede Diskussion über die Notwendigkeit, Menschen mit den so genannten „Kompetenzen des 21. Jahrhunderts" auszustatten, auch eine Diskussion über die Leistungsmessung sein sollte. Den Schwerpunkt der Leistungsmessung auf das, „worauf es ankommt", zu verlagern, ist jedoch nur dann sinnvoll, wenn die Leistungsmessung in der Lage ist, das zu messen, was sie zu messen vorgibt. Die dritte These lautet daher, dass Bewertungen das messen sollten, worauf es ankommt, und sie sollten es *adäquat* messen.

LEISTUNGSMESSUNG IST WICHTIG

Lehrkräfte, Schüler/innen sowie lokale und nationale Entscheidungsträger orientieren sich bei der Festlegung von Lehr- und Lernzielen häufig an den Aufgabentypen, die in lokalen, nationalen und internationalen Prüfungen gestellt werden. Leistungsmessungen signalisieren mehreren Zielgruppen, welche Kenntnisse, Fertigkeiten und Fähigkeiten wichtig sind, und veranschaulichen die Art von Leistung, die der/die Schüler/innen beherrschen sollen. Somit wird das, was wir in Bereichen wie Naturwissenschaften, Mathematik, Lese- und Schreibfähigkeit, Problemlösung, Zusammenarbeit und kritisches Denken bewerten, letztendlich zum Schwerpunkt des Unterrichtsgeschehens. Daher ist es entscheidend, dass unsere Art der Leistungsmessung die Formen von Wissen und Kompetenz sowie die Lernformen, die wir in unseren Klassenräumen stärken möchten, bestmöglich repräsentieren, damit sie positiv im Bildungssystem wirken können.

Aus der Systemperspektive macht es wenig Sinn, massiv in die Reform der Lehrpläne und der Ausbildung von Lehrkräften zu investieren, ohne auch in die Leistungsmessung zu investieren. Lehrpläne, Pädagogik und Leistungsmessung sind eng miteinander verbunden und sollten in gut funktionierenden Bildungssystemen aufeinander abgestimmt sein. Veränderungen in den Lehrplänen und der Pädagogik können durch eine veränderte Ausrichtung der Leistungsmessung und durch die dadurch aufgedeckten Lücken des Bildungssystems vorangetrieben werden, was wiederum die Politikgestaltung und Reformen beeinflusst. Die Fokussierung auf die Leistungsmessung schafft Klarheit über die Lehr- und Lernerwartungen auf den verschiedenen Bildungsebenen und trägt dazu bei, ein gemeinsames Verständnis darüber zu schaffen, was wichtig ist und wie es unterrichtet werden sollte. Die Schlüsselfrage lautet also: Was genau ist wichtig?

VERLAGERUNG DER BILDUNGSZIELE: DER SCHWERPUNKT LIEGT AUF DEN KOMPETENZEN DES 21. JAHRHUNDERTS

Seit mehr als 20 Jahren fordern immer mehr Wirtschaftsführende, Bildungsorganisationen und Forschende eine neue Bildungspolitik, die auf die Entwicklung eines breiten Spektrums von anschluss- und verallgemeinerungsfähigen Fähigkeiten und Kenntnissen abzielt, die oft als „Fähigkeiten des 21. Jahrhunderts" bezeichnet werden (siehe z. B. Pellegrino und Hilton, 2012; Bellanca, 2014). Diesen Forderungen liegt der Gedanke zugrunde, dass der Erfolg in der heutigen globalen Gesellschaft und in einer sich wandelnden Arbeitswelt eine größere Spanne an Fähigkeiten erfordert, die über die traditionellen Lese- und mathematisch-naturwissenschaftlichen Fähigkeiten hinausgehen.

Dabei wird im Wesentlichen davon ausgegangen, dass der Schwerpunkt der Bildung auf der Fähigkeit liegen sollte, (neue) Informationen zu verarbeiten und Probleme zu lösen. Dazu gehört einerseits, über ein fundiertes Fachwissen zu verfügen, andererseits aber auch über ein analytisches, kreatives und kritisches Denkvermögen. Bildung sollte sich darüber hinaus auf umfassendere Fähigkeiten in Bezug auf die eigene Person und andere konzentrieren, wie z. B. soziale und emotionale Fähigkeiten, Toleranz und gegenseitigen Respekt sowie die Fähigkeit zur Selbstregulation und zu einem besseren Verständnis der eigenen Denk- und Lernprozesse.

Gewiss waren diese Fähigkeiten schon immer wichtig. Doch in einer Welt, in der Arbeit durch manuelle und routinemäßige Aufgaben definiert wurde, und in der die sofortige Kommunikation und Informationstechnologien von heute nur Produkte der Vorstellungskraft waren, wurde nur von wenigen Personen erwartet, diese Fähigkeiten zu entwickeln. In den heutigen Wissensökonomien, die sich durch dynamischere und multikulturelle Strukturen auszeichnen, in denen die Bürger/innen sowohl lokal als auch global sich sofort miteinander verständigen und sich selbst organisieren können, werden fortgeschrittene kognitive und sozial-kognitive Kompetenzen als Norm erwartet.

WORUM GEHT ES BEI DEN KOMPETENZEN DES 21. JAHRHUNDERTS?

Bereits vor der Jahrtausendwende hat sich eine wachsende Zahl von Forschenden mit diesem globalen Narrativ befasst und eine Vielzahl internationaler Rahmenwerke ausgearbeitet, worin die Kenntnisse, Fähigkeiten und Einstellungen beschrieben sind, die junge Menschen für die Zukunft benötigen. Innerhalb dieses vergleichsweise dicht besetzten Feldes kommt eine Vielzahl von Begriffen geradezu austauschbar zur Anwendung: „Fähigkeiten/Kompetenzen des 21. Jahrhunderts", „Soft Skills", „interdisziplinäre Fähigkeiten" und „transferierbare Fähigkeiten", um hier nur einige zu nennen. Diese terminologische Zweideutigkeit findet sich auch in der Art und Weise wieder, wie die verschiedenen Rahmenwerke spezifische Kompetenzen definieren (z. B. IKT-Kompetenz vs. digitale Medienkompetenz).

Der Klarheit halber verwendet *Innovating Assessments* mit Bezug auf die in diesen Rahmenwerken erörterte, umfassende Vision von Bildung und die dort beschriebenen Kompetenzen den Begriff „Kompetenzen des 21. Jahrhunderts". Obwohl zwar die Rahmenwerke variieren, werden die Kompetenzen des 21. Jahrhunderts in der Regel wie folgt beschrieben:

- transversal/ (d. h. in vielen Bereichen relevant oder anwendbar);
- multidimensional (d. h. Wissen, Fähigkeiten und Einstellungen umfassend); und
- mit Fähigkeiten und Verhaltensweisen höherer Ordnung verbunden, mit denen Wissen übertragen und komplexe Probleme bewältigt werden können und mit denen man sich an unvorhersehbare Situationen anpassen kann (Voogt und Roblin, 2012).

Über die allgemeine Konvergenz dieser Kernmerkmale hinaus werden die Kompetenzen des 21. Jahrhunderts in den Rahmenwerken auf unterschiedliche Weise ermittelt, organisiert und klassifiziert. Während einige Kompetenzen auf der Grundlage ihrer konzeptionellen Merkmale, z. B. kognitive, interpersonelle und intrapersonelle Kompetenzen (Pellegrino und Hilton, 2012) gruppieren, klassifizieren andere diese eher nach deren Zweck oder Verwendungskontext, z. B. nach „Denkweisen", „Lebensweisen in der Welt", „Arbeitsweisen" und „Arbeitsmitteln" (Binkley et al., 2012).

Abstrahiert man ausgehend von den Besonderheiten der einzelnen Rahmenwerke, so ergeben sich durchgängig einige weitgehend eindeutige Kategorien von Kompetenzen (siehe Abbildung 1). Im Allgemeinen fasst eine Kombination dieser sechs Kategorien das Wesentliche der großen Anzahl von Kompetenzen zusammen, die den verschiedenen Rahmenwerken entstammen, wobei Elemente wie kritisches Denken, kreatives Denken, Kommunikation und IKT-bezogene Kompetenzen sowie die staatsbürgerliche Dimension immer wieder auftauchen. Es ist jedoch zu beachten, dass nicht alle unten aufgeführten allgemeinen Kategorien in sämtlichen Rahmenwerken enthalten sind und auch nicht immer spezifische Kompetenzen denselben allgemeinen Kategorien zugeordnet werden.

Abbildung 1. Allgemeine Kategorien von Kompetenzen des 21. Jahrhunderts

Quelle: Foster (2023), Kapitel 1 in *Innovating Assessments.*

Gemeinsame Kategorien von Kompetenzen des 21. Jahrhunderts zu identifizieren, bietet hilfreiche Einblicke in die Art und Weise, wie sich die allgemeinen Bildungsziele verändern. Dennoch handelt es sich bei diesen Kompetenzen um komplexe Konstrukte. Die Erhebung valider Belege und Interpretationen dessen, wozu Schüler auf der Denk- und Handlungsebene fähig sind, wenn sie diese einsetzen, stellt vor eine Reihe von Herausforderungen. Um die Kompetenzen des 21. Jahrhunderts gut messen zu können, ist es notwendig, innovative Ansätze zur Leistungsmessung sowie neue Erfahrungen in diesem Bereich zu ermöglichen – von der Definition der Konstrukte der Leistungsmessung bis hin zum Aufgabendesign und der Suche nach den richtigen Methoden zur Interpretation der daraus resultierenden Erkenntnisse.

DIE MESSUNG DER KOMPETENZEN DES 21. JAHRHUNDERTS ERFORDERT EIN INNOVATIVES KONZEPT

Das erste Problem bei der Messung von Kompetenzen des 21. Jahrhunderts besteht darin, zu definieren, was gemessen werden soll. Die Kompetenzen sind komplex; sie umfassen mehrere Komponenten, die in der Praxis stark miteinander verwoben sind.

Einerseits erfordert ihr Einsatz eine Kombination von Kenntnissen, Fähigkeiten und Einstellungen – zum Beispiel die Fähigkeit zur Kommunikation. Um effektiv kommunizieren zu können, bedarf es zunächst gewisser Sprachkenntnisse, darüber hinaus jedoch auch eines bestimmten Maßes an schriftlichen, mündlichen oder digitalen Fähigkeiten sowie gewisse Einstellungen gegenüber den Gesprächspartnern. Diese elementaren Bestandteile können in verschiedenen Praxiskontexten auch unterschiedlich ausfallen. Andererseits erfordert der Einsatz einer bestimmten „Art" von Kompetenz im wirklichen Leben oft den gleichzeitigen Einsatz anderer „Kompetenzarten". Das erfolgreiche Lösen eines Problems beispielsweise beinhaltet Aspekte der Metakognition und Selbstregulierung und kann je nach Kontext und Art des Problems auch kreatives Denken und Zusammenarbeit beinhalten. Diese komplexen Zusammenhänge machen es schwierig, Konstrukte in diskrete und unabhängig voneinander messbare Komponenten aufzuschlüsseln wie auch die von den Schüler/innen erbrach-

ten Belege zu isolieren und einer bestimmten Kompetenz zuzuordnen.

Parallel dazu sind die Kompetenzen des 21. Jahrhunderts zumindest teilweise durch Denkprozesse und Verhaltensweisen gekennzeichnet, die über die Fähigkeit zur Reproduktion von Inhaltswissen hinausgehen. Die Fähigkeit, mit neuen Informationen kritisch umzugehen, hängt beispielsweise davon ab, ob man in der Lage ist, zu verstehen, welche zusätzlichen Informationen wie gesucht werden müssen, eine entsprechende Strategie zu planen und auszuführen sowie bei der Lösung der Aufgabe durchzuhalten und/oder zu entscheiden, wer um Hilfe oder Feedback gebeten werden kann. Diese Verhaltens- und Denkweisen müssen in Leistungsmessungen sichtbar gemacht werden, wenn damit eine Aussage über die Kompetenz von Schüler/innen getroffen werden soll. In Bezug auf eine große Anzahl der Kompetenzen des 21. Jahrhunderts bedeutet dies, dass Testumgebungen geschaffen werden müssen, die den Schüler/innen Handlungswerkzeuge sowie Wahlmöglichkeiten und Gelegenheiten zur Erkundung und Iteration ihrer Ideen bieten. Dies erfordert, dass die Bewertungsaufgaben und -merkmale über die statischen, geschlossenen Antwortarten hinausgehen, die typischerweise in groß angelegten Leistungsmessungen verwendet werden, um einen reichhaltigeren Satz von Daten darüber zu generieren, wie Schüler/innen denken und handeln.

Die Entwicklung einer nächsten Generation von Einschätzungen von Wissen und Fähigkeiten, die dieser Vision der Bildung des 21. Jahrhunderts entsprechen, bringt daher eine Reihe von Herausforderungen mit sich, die von jenen, die die Leistungsmessungen erstellen, bewältigt werden müssen. Dazu gehören die Definition der Zielkonstrukte der Bewertung, die Identifikation der relevanten Situationen, in denen diese beobachtet werden können, die Replikation ihrer Kernmerkmale in Bewertungsumgebungen, die Übersetzung von Spuren von Handlungen innerhalb dieser Umgebungen in Beweise sowie die Entwicklung geeigneter Modelle zur Interpretation und Bewertung der Belege, um belastbare Aussagen über die Leistung zu treffen.

Die folgenden Abschnitte stützen sich auf die Kernaussagen und die fortschrittlichsten Praxisbeispiele in der OECD-Publikation *Innovating Assessments* und beleuchten den Weg, der beim Entwickeln von Leistungsmessungen eingeschlagen werden muss – einschließlich der wichtigsten Entscheidungen, die in Betracht gezogen werden müssen, und der neuen Instrumente, die auf diesem Weg helfen können. Am Ende des vorliegenden Dokuments finden sich einige Überlegungen zu der Rolle, die Bildungsbehörden zusammen mit anderen Stakeholdern als Teil eines breiteren Rahmens der internationalen Zusammenarbeit spielen können, um die Agenda der „Leistungsmessungen der nächsten Generation" voranzubringen.

LEISTUNGSMESSUNGEN DER NÄCHSTEN GENERATION: DESIGNPRINZIPIEN UND BEISPIELE

Die Messung von Bildungsergebnissen ist nicht so einfach wie die Messung von Größe oder Gewicht. Leistungsmessungen gewähren keinen direkten Einblick in das Gehirn eines Schülers oder einer Schülerin; die zu messenden Attribute sind geistiger Natur, die nach außen hin nicht sichtbar sind. Eine Leistungsmessung ist also ein Instrument, das dazu dient, das Verhalten der Schüler/innen zu beobachten und Daten hervorzubringen, die Rückschlüsse auf ihre Kenntnisse und Fähigkeiten zulassen. Zu entscheiden, was und wie gemessen werden soll, ist nicht so einfach, wie es den Anschein haben könnte. Dies gilt umso mehr, wenn die Ziele der Leistungsmessung komplexe Konstrukte und Leistungen sind.

Innovating Assessments stellt Schlüsselideen für die Entwicklung der nächsten Generation von Leistungsmessungen vor, die die von den Schüler/innen benötigten Kompetenzen messen, und liefert umsetzbare Informationen für Entwickler/innen von Leistungsmessungen, Pädagog/innen und politische Entscheidungsträger. Die Messung *dessen, worauf es ankommt*, erfordert Innovationen in allen Phasen des Testdesigns – von dem, was wir messen, bis zu der Art und Weise, wie wir es tun. Die *adäquate* Messung dessen, worauf es ankommt, erfordert einen prinzipiengeleiteten Entwurfsprozess und die Nutzung digitaler Technologien, um relevante Erkenntnisse über die Kompetenzen der Schüler/innen zu gewinnen und innovative Analysemethoden anzuwenden, um diese Erkenntnisse sinnvoll zu nutzen.

BEURTEILUNG ALS EIN PROZESS DES SCHLUSSFOLGERNS AUS BELEGEN

Der Prozess, der Rückschlüsse auf das Wissen und die Fähigkeiten der Schüler/innen zulässt, ist eine Argumentationskette, die sich aus den Belegen der Kompetenzen der Schüler/innen ergibt und die für alle Leistungsmessungen kennzeichnend ist, von Klassenarbeiten und standardisierten Leistungstests über computergestützte Nachhilfeprogramme bis hin zu den Gesprächen, die Schüler/innen mit ihren Lehrpersonen führen, während sie ein mathematisches Problem lösen oder die Bedeutung eines Textes diskutieren. Die erste Frage im Beurteilungsprozess lautet: „Belege wofür?" Daten liefern ihre Deutung nicht zugleich mit, ihr Wert als Beleg entsteht vielmehr erst durch einen Interpretationsrahmen. Beurteilungen liefern Daten wie schriftliche Aufsätze, Markierungen auf Antwortbögen, Präsentationen von Projekten oder die Erklärungen der Schüler/innen zu ihrem Weg der Problemlösung, jedoch werden diese Daten nur im Hinblick auf Vermutungen darüber, wie Schüler/innen Wissen und Fähigkeiten erwerben, zu Belegen.

Pellegrino et al. (2001) beschreiben diesen Prozess der Inferenz anhand

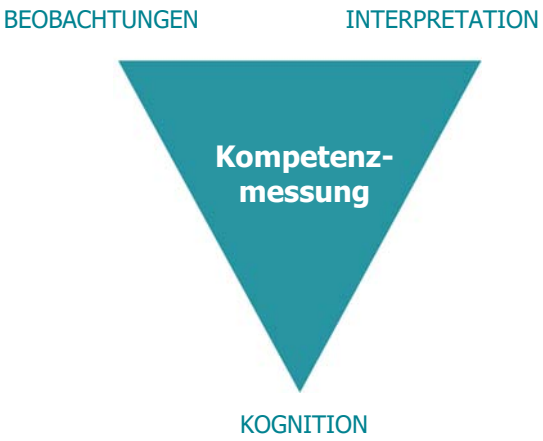
von Belegen als einen Dreiklang aus drei miteinander verknüpften Komponenten: das Assessment Triangle (siehe Abbildung 2). Die Eckpunkte des Dreiecks stellen die drei Schlüsselkomponenten dar, die jeder Kompetenzmessung zugrunde liegen: ein Modell der kognitiven Fähigkeiten und des Lernens der Schüler/innen im bewerteten Bereich; eine Reihe von Annahmen und Grundsätzen über die Art der Beobachtungen, die einen Beleg für die Kompetenzen der Schüler/innen liefern; und ein Interpretationsverfahren, um den Beleg unter Berücksichtigung des Bewertungszwecks und des Verständnisses der Schüler/innen sinnvoll zu gestalten. Diese drei Komponenten können explizit oder implizit sein, aber eine Kompetenzmessung kann nicht entworfen und durchgeführt oder evaluiert werden, ohne jede einzelne Komponente zu berücksichtigen. Die drei Komponenten werden als Eckpunkte eines Dreiecks dargestellt, da jede mit den beiden anderen verbunden und von ihnen abhängig ist. Das Assessment Triangle bietet einen hilfreichen Rahmen, um die Grundlagen aktueller Kompetenzmessungen zu analysieren und festzumachen, wie gut sie die angestrebten Ziele erreichen, sowie für die Entwicklung zukünftiger Kompetenzmessungen und die Feststellung ihrer Validität (z. B. Pellegrino, et al., 2016).

Abbildung 2. Das Assessment Triangle

KOGNITION
Theorien, Modelle und Daten darüber, wie Schüler/innen Wissen darstellen und Kompetenzen in einem Lehr- und Lernbereich entwickeln.

BEOBACHTUNGEN
Aufgaben oder Situationen, die es ermöglichen, die Leistungen von Schülern zu beobachten.

INTERPRETATION
Methoden zur Auswertung der von den Schülern erbrachten Leistungen.



Quelle: Pellegrino et al. (2001).

Die *Kognitions*komponente des Dreiecks bezieht sich auf Theorie, Daten und eine Reihe von Annahmen darüber, wie die Schüler/innen Wissen repräsentieren und Kompetenzen in einem intellektuellen Bereich entwickeln (z. B. Brüche; Newton'sche Gesetze; Thermodynamik). Für jede Bewertung wird eine Theorie der Kompetenz in dem Bereich benötigt, um zu bestimmen, welches Wissen und welche Fähigkeiten für den beabsichtigten Anwendungskontext wichtig sind – sei es, um die Kompetenzen zu charakterisieren, die die Schüler/innen zu einem bestimmten Zeitpunkt erworben haben, um eine summative Leistungsüberprüfung vorzunehmen, oder um formative Leistungsüberprüfungen vorzunehmen, um den Unterricht anschließend so zu gestalten, dass der größtmögliche Lernerfolg erzielt wird. Eine zentrale Prämisse ist, dass die kognitive Theorie das wissenschaftlich glaubwürdigste Verständnis der typischen Art und Weise darstellt, in der Lernende Wissen repräsentieren und Fachwissen im jeweiligen Schwerpunktbereich entwickeln.

Jede Leistungsmessung basiert auch auf einer Reihe von Annahmen und Grundsätzen über die Art der Aufgaben oder Situationen, die die Schüler/innen dazu veranlassen, etwas zu sagen, zu tun oder zu schaffen, das wichtige Kenntnisse und Fähigkeiten demonstriert. Die Aufgaben, die Schüler/innen im Rahmen eines Tests lösen sollen, müssen

sorgfältig konzipiert sein, um Belege zu generieren, die mit der kognitiven Lerntheorie verknüpft sind, und um jene Inferenzen und Entscheidungen zu stützen, die auf der Grundlage der Testergebnisse getroffen werden. Die *Beobachtungs*komponente des Assessment Triangle stellt eine Beschreibung oder eine Reihe von Spezifikationen für Testaufgaben dar, die aufschlussreiche Antworten von den Schüler/innen hervorrufen sollen. Bei der Leistungsmessung hat man die Möglichkeit, einen kleinen Ausschnitt der Welt zu strukturieren, um Beobachtungen zu machen. Beim Design der Leistungsmessung kann dies genutzt werden, um den Wert der gesammelten Daten durch die Linse der zugrundeliegenden Annahmen darüber, wie Schüler/innen in dem Bereich lernen, zu maximieren.

Leistungsmessungen erfordern auch bestimmte Annahmen und Modelle, um die aus Beobachtungen gewonnenen Erkenntnisse zu interpretieren. Die *Interpretations*komponente des Dreiecks umfasst alle Methoden und Werkzeuge, die verwendet werden, um aus fehlbaren Beobachtungen Schlüsse zu ziehen. Sie drückt aus, wie die aus einer Reihe von Testaufgaben abgeleiteten Beobachtungen Belege für die zu bewertenden Kenntnisse und Fähigkeiten darstellen. Im Kontext von groß angelegten Kompetenzmessungen ist die Interpretationsmethode in der Regel ein statistisches Modell, d. h. eine Charakterisierung oder Zusammenfassung von Mustern, die man bei unterschiedlichen Kompetenzniveaus der Schüler/innen in den Daten erwarten würde. Bei der Kompetenzmessung im Unterricht wird die Interpretation oft weniger formell von der Lehrkraft vorgenommen und basiert eher auf einem intuitiven oder qualitativen als auf einem formalen statistischen Modell. Selbst informell treffen die Lehrpersonen jedoch koordinierte Entscheidungen darüber, welche Aspekte des Verständnisses und des Lernens der Schüler/innen relevant sind, wie ein/e Schüler/in eine oder mehrere Aufgaben gelöst hat und was die Leistungen über das Wissen und das Verständnis der betreffenden Person aussagen.

Ein entscheidender Punkt ist, dass jede der drei Komponenten des Assessment Triangle nicht nur für sich allein Sinn ergeben muss, sondern auch mit beiden anderen Komponenten auf sinnvolle Weise in Verbindung stehen muss, um zu einer effektiven Beurteilung und zu fundierten Inferenzen zu führen. Um eine gültige und wirksame Kompetenzmessung zu erhalten, müssen also alle drei Eckpunkte des Dreiecks synchron zusammenarbeiten. In Anerkennung der Tatsache, dass Leistungsmessung ein evidenzbasierter Denkprozess ist, hat es sich als nützlich erwiesen, den Entwicklungsprozess von Leistungsmessungen systemisch als evidenzbasierten Designprozess (Evidence-Centered Design, ECD) zu gestalten (z. B. Mislevy und Haertel, 2006; Mislevy und Riconscente, 2006) – siehe Abbildung 3 für einen Überblick über die verschiedenen Komponenten des ECD-Modells.

Abbildung 3. Testerstellung als evidenzbasierter Designprozess

Phasen der Festlegung des konzeptionellen Rahmens einer Leistungsmessung

FESTLEGUNG VON ZIELEN UND SCHWERPUNKTEN

FESTLEGUNG DER TESTDOMÄNE

- Sammeln von Informationen über die Domäne (Domänenanalyse), einschließlich ihrer Hauptkomponenten und des Spektrums an Problemen und Situationen, in denen die angestrebten Kenntnisse und Fähigkeiten Anwendung finden.
- Domänenmodellierung: Festlegung von Bewertungsansprüchen (was gemessen werden soll), Daten (wie es gemessen wird) und Begründungen (warum der gewählte Messansatz angemessen ist).

OPERATIONALISIERUNG DER BEWERTUNG (ECD-RAHMEN)

FESTLEGUNG, WIE DIE SCHÜLERLEISTUNGEN IN DEM BEREICH AUSSEHEN (DAS SCHÜLERMODELL)

Definition der Variablen (Wissen, Fähigkeiten und Einstellungen), über die Aussagen gemacht werden sollen, der Beziehungen zwischen diesen Variablen und ob diese Variablen dynamisch sind (wenn ein gewisses Lernen erwartet wird).

- Bereitstellung einer detaillierten Vision dessen, was die Schüler/innen auf den verschiedenen Leistungsniveaus verstehen und können, von den niedrigsten bis zu den höheren Niveaus der Beherrschung jeder Variablen.

FESTLEGUNG DER SITUATIONEN, IN DENEN EIN LEISTUNGSNACHWEIS ERBRACHT WERDEN KANN (DAS AUFGABENMODELL)

- Spezifizierung der Aufgaben, bei denen die Testteilnehmenden ihre Fähigkeiten unter Beweis stellen können, wie z. B. bei vordefinierten Fragen oder Aufgaben (z. B. Multiple-Choice-Aufgaben, Aufgaben zur Neuordnung oder Vervollständigung) oder in Umgebungen, in denen die Situation durch die Handlungen der Testteilnehmenden gestaltet wird (z. B. Simulationen, Spiele).
- Festlegung der Faktoren, die die Komplexität und das Wissen beeinflussen, und der in die Aufgabe eingebetteten Ressourcen, einschließlich Feedback oder Hilfsmittel zur Erleichterung des Lernens (wenn Lernen erwartet wird).

FESTLEGUNG VON LEISTUNGSKENNZAHLEN UND INDIKATOREN (DAS EVIDENZMODELL)

- Definition der Evidenzregeln: Zuweisung einer Punktzahl oder eines Wertes zu dem, was Testteilnehmende tun (z. B. Fragen richtig/falsch beantworten, bestimmte Handlungen/Entscheidungen in einer bestimmten Situation treffen).
- Aufbau eines statistischen Modells, das Daten über Aufgaben hinweg in Bezug auf aktualisierte Überzeugungen über Variablen des Schülermodells zusammenfasst.

Quelle: Piacentini (2023), Kapitel 6 in *Innovating Assessments.*

ERNEUERUNG DER *KOGNITIONS*KOMPONENTE: DEFINITION VON KONSTRUKTEN DER KOMPETENZMESSUNG

Bei der Entwicklung von Kompetenzmessungen kommt der klaren Definition der Zieldomäne und der Beschreibung der Kenntnisse, Fähigkeiten, Einstellungen und Anwendungskontexte, die der Leistung in dieser Domäne zugrunde liegen, die höchste Bedeutung zu. Ist die Domäne nicht klar definiert, kann weder die Sorgfalt, die an anderer Stelle für die Testentwicklung aufgewendet wird, noch eine komplexe psychometrische Analyse nach der Datenerhebung diese Unzulänglichkeit ausgleichen (Mislevy und Riconscente, 2006). Es ist weitaus wahrscheinlicher, dass eine Leistungsmessung ihren Zweck erfüllt, wenn die Art des Konstrukts das Design der relevanten Aufgaben sowie die Entwicklung von konstruktbasierten Auswertungskriterien und Bewertungsrastern leitet (Messick, 1994).

Wie bereits erörtert, wird diese entscheidende Tätigkeit mit zunehmender Komplexität der Domäne und der Zielkonstruktion(en) immer schwieriger. Die Arten von Problemen oder Lernaktivitäten, die die Kompetenzen des 21. Jahrhunderts erfordern und fördern, verlangen eine andere Kombination von Wissen, Fähigkeiten und Einstellungen, und der Anwendungskontext spielt eindeutig eine Rolle, um zu bestimmen, welche dieser Elemente am wichtigsten sind und wie genau sie zum Ausdruck gebracht werden können. Das bedeutet, dass es wichtig ist, bereits in den ersten Phasen der Testentwicklung deutlich zu machen, was von den Schüler/innen erwartet wird, dass die Schüler/innen durch ihre Leistung im Text demonstrieren sollen.

FRÜHZEITIGE ENTSCHEIDUNGEN ÜBER DEN SCHWERPUNKT DER MESSUNG DER KOMPETENZEN DES 21. JAHRHUNDERTS

Bei der Entscheidung darüber, was gemessen werden soll, ist die Auswahl eines Rahmens oder einer Liste von Kompetenzen des 21. Jahrhunderts und die Entwicklung eines einzigen Messinstruments für jede beschriebene Kompetenz möglicherweise nicht die beste Wahl. Da die Kompetenzen des 21. Jahrhunderts multidimensional und in der Praxis stark miteinander verbunden sind, könnte eine produktivere Strategie darin bestehen, zu messen, wie Schüler/innen Wissen schaffen und verschiedene Arten komplexer Probleme lösen, allein oder in Zusammenarbeit in verschiedenen Anwendungskontexten. Wenn wir uns ein Bild davon machen, was die Schüler/innen in offenen, erweiterten Problemlösungsaktivitäten tun, erhalten wir Informationen darüber, wie sie in authentischeren Szenarien die vielfältigen Kompetenzen des 21. Jahrhunderts zur Anwendung bringen.

Wie in Abbildung 4 dargestellt, können drei miteinander verknüpfte Fragen besonders hilfreiche Anhaltspunkte für die Festlegung des Schwerpunkts der nächsten Generation von Leistungsmessungen bieten:

Abbildung 4. Anfängliche Entscheidungen in Bezug auf den Fokus der Leistungsmessungen des 21. Jahrhunderts

Quelle: Piacentini und Foster (2023), Kapitel 3 in *Innovating Assessments.*

- In welchen praktischen Kontexten können sich Schüler/innen an Messaktivitäten beteiligen? Diese Entscheidung bezieht sich auf die Anerkennung der Kenntnisse, Fähigkeiten und Einstellungen, die Schüler/innen für eine bestimmte Art von Aktivität in einem bestimmten Praxiskontext benötigen (d. h. die Aktivität innerhalb der Grenzen eines Fachgebiets zu verorten oder sie fächerübergreifend zu gestalten und den Anwendungskontext zu spezifizieren).

- Welche Arten von Leistungen und damit verbundenen Aktivitäten sollen in Bezug auf den aktuellen Stand der Schüler/innen untersucht werden? Diese Entscheidung bezieht sich auf die explizite Definition der Messaktivitäten und der relevanten Vorgangsweisen, die die Schüler/innen während der Durchführung dieser Aktivitäten zeigen sollen.

- Wird die Bewertung als Einzel- oder als Gruppenaktivität organisiert? Diese Entscheidung bezieht sich auf die ausdrückliche Festlegung, ob, wann und zu welchem Zweck ein Test den Schüler/innen die Möglichkeit zur Interaktion mit anderen – realen oder simulierten – Akteuren bieten kann.

Relevante Aktivitäten zur Bewertung der Kompetenzen des 21. Jahrhunderts

Das Lösen komplexer Probleme erfordert eine Vielzahl kognitiver, metakognitiver, einstellungsbezogener und sozio-emotionaler Fähigkeiten. Jedoch sind nicht alle Testaufgaben geeignet, einen so reichhaltigen Fundus an Erkenntnissen über die Lernenden zu liefern. Traditionelle Modelle des Problemlösens, die als Phasenmodelle bekannt sind (z. B. Bransford und Stein, 1984), gehen davon aus, dass alle Probleme gelöst werden können, wenn folgende Phasen durchlaufen werden: (1) Identifizieren des Problems; (2) Entwickeln alternativer Lösungen; (3) Bewertung dieser Lösungen; (4) Umsetzung der gewählten Lösung; und (5) Bewertung der Wirksamkeit der Lösung. Obwohl diese Be-

schreibungen allgemeiner Prozesse nützlich sind, könnten sie fälschlicherweise den Eindruck erwecken, dass Problemlösen eine einheitliche Tätigkeit ist (Jonassen, 1992). In der Realität ist es jedoch so, dass sich Probleme in vielerlei Hinsicht unterscheiden, u. a. durch den Kontext, in dem sie auftreten, durch den Grad ihrer Struktur oder Offenheit und durch die Kombination von Fähigkeiten, die eingesetzt werden müssen, um eine Lösung zu finden.

Es gibt eine Vielzahl von Problemstellungen und Aktivitäten, die Schüler/innen vorgelegt werden können, um die Kompetenzen des 21. Jahrhunderts zu messen. Zu den Gruppen von Aufgaben, die wahrscheinlich valide Belege darüber liefern, ob die Lernerfahrungen die Schüler/innen auf ihre Zukunft vorbereitet haben, gehören zum Beispiel: (1) Suche, Bewertung und Austausch von Informationen; (2) Verständnis, Modellierung und Optimierung von Systemen; und (3) Entwurf kreativer Produkte. Dies ist keine erschöpfende Typologie von Aufgaben; die Arten von Problemen und Aktivitäten, auf die die Schüler/innen vorbereitet werden müssen, entwickeln sich weiter. Außerdem schließen sich diese drei Gruppen nicht gegenseitig aus, sondern überschneiden sich vielmehr bis zu einem gewissen Grad. Sie illustrieren jedoch Problemtypen, die ganz unterschiedliche Kompetenzen und damit verbundene Kenntnisse, Fähigkeiten und Einstellungen erfordern. Kasten 1 enthält einige Beispiele dafür, wie die nächste Generation von Leistungsmessungen der ersten Aufgabengruppe aussehen könnte.

KASTEN 1.

RELEVANTE AKTIVITÄTEN FÜR DIE BEWERTUNG DER KOMPETENZEN DES 21. JAHRHUNDERTS

Informationssuche, -auswertung und -weitergabe

Bei dieser Art von Aktivitäten besteht das zu lösende Hauptproblem bzw. Lernziel darin, Informationen zu suchen und zu nutzen, um eine belastbare Inferenz zu ziehen. Die Abfolge der Aufgaben eines Tests sollte die Schüler/innen dazu anregen, ihren Informationsbedarf zu ermitteln, Informationsquellen online oder offline zu finden, Informationen aus den Quellen zu extrahieren, zu organisieren und zu vergleichen, Informationskonflikte zu lösen und Entscheidungen darüber zu treffen, welche Informationen wie weitergegeben werden sollen. Diese Aktivitäten werden häufig als Problemlösen durch Informationsverarbeitung definiert (Brand-Gruwel et al., 2005; Wolf et al., 2003). Untersuchungen zeigen, dass viele Schüler/innen nicht in der Lage sind, Informationsprobleme erfolgreich zu lösen (Bilal, 2000; Large und Beheshti, 2000). Diese konzentrieren sich darauf, wie Schüler/innen mit verschiedenen Arten von Medien interagieren, und können auf praktisch jeden Wissensbereich (d. h. auf den praktischen Kontext) angewendet werden. Sie betonen kritisches Denken, Synthese und Argumentation, verantwortungsvolle Kommunikation und selbstgesteuerte Lernfähigkeiten als Kernkompetenzen.

Es gibt mehrere Beispiele für Leistungsmessungen, die sich auf Informationsprobleme konzentrieren. In einigen Fällen ist der Test vollständig in eine Lernerfahrung integriert, und die Belege werden „heimlich" durch die Analyse der Sequenzen von Entscheidungen, die die Schüler/innen treffen, und anhand des Ergebnisses ihrer Informationssuche extrahiert. In der Betty's-Brain-Umgebung (Biswas, 2015) unterrichten die Schüler/innen beispielsweise eine virtuelle Agentin, Betty, über ein wissenschaftliches Phänomen. Dazu durchsuchen sie mit Hyperlinks versehene Ressourcen und erstellen eine Concept Map, die ihr entstehendes Verständnis des Phänomens darstellt. Die Schüler/innen können Betty bitten, sich einem Test zu stellen, in welchem sie anhand der in der Concept Map dargestellten Informationen antwortet; Bettys Leistung in diesem Test gibt den Schüler/innen Aufschluss über falsche oder fehlende Elemente in der Map.

In anderen Beispielen werden Tools zur Informationssuche und -verwaltung in virtuelle Welten eingebettet. Im Rahmen des NAEP-SAIL-Projekts „Virtual World for Online Inquiry" (Coiro et al., 2019) wurde eine virtuelle Plattform entwickelt, die eine Kleinstadt simuliert, in der Schüler/innen vor eine offene Lernaufgabe gestellt werden (z. B. zu ermitteln, ob ein historisches Artefakt im örtlichen Museum ausgestellt werden sollte) und ihr Wissen aufbauen, indem sie eine Untersuchungsstrategie mit einem virtuellen Partner planen, Fragen an virtuelle Experten stellen, nach Informationen in einer Web-Umgebung oder in einer virtuellen Bibliothek suchen und dabei verschiedene digitale Werkzeuge verwenden, um Notizen zu machen und einen Bericht zu redigieren. Die Umgebung umfasst adaptive Funktionen wie Hinweise, Aufforderungen und Niveaustufen, die den Schüler/innen helfen, ihre Rechercheprozesse zu regulieren und eine effiziente und effektive Informationsbeschaffung fördern.

Andere interessante Beispiele beziehen sich auf die Fähigkeiten der Schüler/innen zur Überprüfung von Fakten und zum Informationsaustausch in offenen vernetzten Umgebungen. Spiele wie „Fake It To Make It" (Urban et al., 2018), „Bad News" (Roozenbeek und van der Linden, 2019) oder „Go Viral!" (Basol et al., 2021) bringen den Spielenden gängige Techniken zur Verbreitung von Fehlinformationen bei, in der Hoffnung, dass sie dadurch darauf vorbereitet werden, auf diese zu reagieren. In „The Misinformation Game" können sich die Teilnehmenden mit Social-Media-Beiträgen auseinandersetzen, indem sie aus Optionen wie „Gefällt mir", „Gefällt mir nicht" „Teilen", „Markieren" und „Kommentieren" wählen, und sie erhalten dynamisches Feedback (d. h. Änderungen ihrer eigenen simulierten Followerzahl und ihres Glaubwürdigkeitswerts), je nachdem, wie sie mit zuverlässigen oder unzuverlässigen Informationen interagieren (van der Linden et al., 2020).

Quelle: Piacentini und Foster (2023), Kapitel 3 in *Innovating Assessments*.

Praxiskontexte oder Anwendungsbereiche

Die Kompetenzen des 21. Jahrhunderts werden zwar weithin als transversal oder interdisziplinär betrachtet, was es jedoch bedeutet, Probleme zu lösen, kritisch zu denken oder kreativ zu sein, kann in verschiedenen Kontexten ganz unterschiedlich aussehen. Diese Fähigkeiten werden weder in einem Vakuum eingesetzt noch beobachtet, und sie können kaum domänenneutral bewertet werden. Daher sollte bei der Festlegung des Schwerpunkts einer Leistungsmessung die Rolle und Bedeutung des domänenspezifischen Wissens von Anfang an deutlich gemacht werden. Im Kontext einer Leistungsmessung werden die Fähigkeiten der Schüler/innen immer in einem bestimmten Kontext oder einer bestimmten Situation beobachtet, und ihr Wissen über diesen Kontext oder diese Situation beeinflusst die Art der Strategien, die sie anwenden, sowie das, was sie zu leisten vermögen. Der Versuch, völlig dekontextualisierte Probleme oder Szenarien zu entwerfen, gefährdet auch die Validität: Wenn kein Wissen benötigt wird, um eine Aufgabe zu lösen, kann ein Test dann wirklich behaupten, die Arten von komplexen Problemlösungskompetenzen zu messen, an denen er angeblich interessiert ist?

Die Leistungsmessungen der nächsten Generation können in einem bestimmten Wissensbereich kontextualisiert werden oder mehrere Disziplinen umfassen. Fächerübergreifend bedeutet hier nicht domänenübergreifend, da die Kompetenzen, die die Schüler/innen bei fächerübergreifenden Aufgaben zeigen, immer noch von einem genau definierten Wissensbestand abhängen; nur ist dieses Wissen nicht von den Grenzen eines einzelnen Fachbereichs beschränkt. Die am weitesten verbreiteten Leistungsmessungen von Lernergebnissen beziehen sich auf eine einzige Disziplin (z. B. Mathematik, Biologie, Geschichte) und konzentrieren sich auf die Reproduktion der erworbenen Kenntnisse und Vorgehensweisen, die für diese relevant sind. Wird eine Messung der Kompetenzen des 21. Jahrhunderts im Kontext einer Fachdomäne in den Blick genommen, könnten neue Tests ein besseres Gleichgewicht zwischen der Messung von Fachwissen und der Messung der Fähigkeit der Schüler, dieses Wissen in authentischen Kontexten und auf neue Probleme anzuwenden, herstellen. Die Tests könnten die Schüler/innen zu Vorgangsweisen auffordern, die widerspiegeln, wie Fachwissen zur Lösung beruflicher und alltäglicher Probleme eingesetzt

wird. In Geschichte könnten die Schüler/innen beispielsweise aufgefordert werden, gemeinsam eine historische Darstellung eines Ereignisses zu untersuchen und Verzerrungen zu finden. In den Naturwissenschaften könnten die Schüler/innen aufgefordert werden, ein wissenschaftliches Phänomen in einem virtuellen Labor zu erforschen, wobei sie einschlägige Hilfsmittel verwenden und die Abfolge von Entscheidungen durchlaufen, die echte Wissenschaftler/innen in ihrer beruflichen Praxis treffen (siehe Kasten 2 für ein ausführlicheres Beispiel).

KASTEN 2.

DOMÄNENSPEZIFISCHE MESSUNGEN KOMPLEXER FÄHIGKEITEN

Überprüfung der Entscheidungsfindung von Schüler/innen in den Bereichen Wissenschaft und Technik

Komplexes Problemlösen, insbesondere in den Bereichen Wissenschaft und Technik, sind eine Kernkompetenz der modernen Welt, und viele neuere Wissenschaftsstandards stellen diese Kompetenz in den Mittelpunkt. Allerdings erfassen die Leistungsmessungen in der Regel nicht die Schlüsselprozesse und -entscheidungen, die das Problemlösen im wirklichen Leben mit sich bringt, und sind daher nur begrenzt geeignet, um aussagekräftige Rückschlüsse auf die Kompetenzen der Schüler/innen zu ziehen.

Das Lösen von Aufgaben, die typischerweise in Schulprüfungen und Lehrbüchern vorkommen, erfordert das Erkennen und Befolgen eines einzigen, gut etablierten Vorgehens. Diese Aufgaben können insofern kompliziert sein, als sie mehrere Schritte erfordern, aber es sind nur sehr wenige Entscheidungen erforderlich – entweder man kennt das richtige Vorgehen oder nicht. Dies ist nicht die Art und Weise, wie komplexe Probleme gelöst werden. Erfahrene Wissenschaftler und Ingenieure sind keine Expert/innen, weil sie gut darin sind, ein bestimmtes Vorgehen abzuarbeiten oder eine bestimmte Methode anzuwenden, sondern weil sie gut darin sind, ihr Wissen und ihre fachlichen Kompetenzen anzuwenden, um Probleme zu lösen, für die es keine vollständigen Informationen und keine vordefinierten Lösungsschritte gibt. Im Gegensatz zu Schulaufgaben enthalten reale Aufgaben eine Mischung aus relevanten und irrelevanten Informationen, und einige der anspruchsvollsten Aspekte bei ihrer Lösung beziehen sich auf die Beantwortung von Fragen wie „Welche Informationen werden benötigt?", „Welche Konzepte sind relevant?", „Was ist ein guter Plan?", „Welche Inferenzen sind durch die Belege gerechtfertigt?".

Laut Wieman und Price (2023) sollten schulische Aufgaben (und damit auch Aufgaben im Rahmen von Leistungsmessungen) mehr wie authentische Probleme aussehen: Sie sollten den Schüler/innen die Möglichkeit geben, sich mit der Art von Entscheidungsfindung zu beschäftigen und sie zu üben, mit der Praktiker in der realen Welt konfrontiert sind, d. h. zu lernen, wie eine Wissenschaftlerin oder ein Ingenieur zu denken und zu argumentieren. Ein Problem kann authentisch sein, wenn Entscheidungen getroffen werden müssen, anstatt dass einer vorgeschriebenen Vorgangsweise gefolgt wird, und auf das Wissen beschränkt sein, das von den Schüler/innen auf einem bestimmten Niveau erwartet wird. Der Schlüssel dazu ist ein gutes Verständnis der Entscheidungen, mit denen Praktiker konfrontiert sind (Kognitionskomponente), und die Anwendung dieses Wissens bei Aufgabendesign und Bewertungsmethoden.

Um das richtige Gleichgewicht zwischen Authentizität und Praktikabilität bei Leistungsmessungen zu finden, müssen Aufgaben und Fragen ausgewählt werden, die die Problemlösenden in angemessenem Umfang einschränken. Zu viel Einschränkung bedeutet, dass wichtige Ressourcen und Entscheidungsprozesse nicht erforscht werden, während zu wenig Einschränkung zu Antworten führt, die so stark variieren können und es dadurch unmöglich machen, die Stärken und Schwächen der Testteilnehmenden im Einzelnen zu bewerten und zu vergleichen.

Quelle: Wieman und Price (2023), Kapitel 4 in *Innovating Assessments*.

Während der Einbezug von Wahlmöglichkeiten und authentischen Problemen in fachliche Leistungsmessungen wichtige Wege für die Erneuerung aktueller Messpraktiken darstellt, könnte es auch ein wertvoller Ansatz sein, Leistungsmessungen der nächsten Generation über mehrere Bereiche hinweg durchzuführen. Eine Möglichkeit, Schüler/innen in fächerübergreifende Aufgaben einzubinden, könnte darin bestehen, Testsituationen vorzuschlagen, in denen sie als verantwortungsbewusste Bürger/innen handeln müssen, indem sie sich mit Problemen auseinandersetzen, die eine Gruppe von Gleichaltrigen, eine Nachbarschaft oder eine größere Gemeinschaft betreffen. Moderne simulationsbasierte Tests können viele dieser erfahrungsbasierten Lernsituationen einbeziehen und bieten die Möglichkeit, soziale Entscheidungen zu treffen und ein empathisches Verständnis zu entwickeln, indem man sich selbst durch einen Avatar darstellt (Raphael et al., 2009). Diese Kontexte können besonders geeignet sein, um sozio-emotionale Fähigkeiten wie Kommunikation, Kooperation, Emotionsregulation und Empathie zu messen. Rollenspiele wurden vermehrt entwickelt, um diese Fähigkeiten auf verdeckte Weise zu messen, wie z. B. „Hall of Heroes" (Irava et al., 2019). Eine große Herausforderung bei der Entwicklung fächerübergreifender Kompetenzmessungen besteht jedoch darin, dass es in diesen „Domänen" an soliden Theorien über die Entwicklung von Wissen und Fähigkeiten fehlt. Genau zu definieren, welche Faktoren konstruktrelevant oder -irrelevant sind, und was eine „gute Leistung" kulturübergreifend gültig ausmacht, sind ähnliche Herausforderungen.

Individuelle vs. kollaborative Aufgaben

Gruppenarbeit wird weltweit zunehmend als pädagogische Praxis eingesetzt, obwohl es für Lehrkräfte eine Herausforderung ist, kollaboratives Lernen effektiv zu strukturieren und zu moderieren (Gillies, 2016). Forschende und Lehrkräfte sind sich zunehmend der positiven Auswirkungen bewusst, die die Zusammenarbeit auf die Lernfähigkeit der Schüler/innen haben kann. Die Forschung zeigt, dass gemeinschaftliches Arbeiten sowohl die akademischen Leistungen als auch die Sozialisationsfähigkeiten fördert, wobei diese positiven Auswirkungen für alle Altersgruppen und Fächer gelten (Baines et al., 2007; Gillies und Boyle, 2010). Praktiken der formativen Leistungsmessung sind diesem Trend gefolgt, da immer mehr Lehrkräfte auf der ganzen Welt Bewertungsraster anwenden, um die Fähigkeit ihrer Schüler/innen, in Gruppen zu arbeiten, zu bewerten. Bei den summativen Leistungsmessungen sind die Fortschritte viel zögerlicher, wenngleich es

auch bemerkenswerte Ausnahmen gibt (siehe Kasten 3 für zwei Beispiele in groß angelegten Leistungsmessungen).

KASTEN 3.

MESSUNG DER ZUSAMMENARBEIT VON SCHÜLER/INNEN BEI UMFANGREICHEN TESTS

PISA 2015 und Assessment and Teaching of 21st Century Skills (ATC21S)

Im Rahmen der PISA-Domäne des kollaborativen Problemlösens bilden drei Kompetenzen den Kern der Kollaborationsdimension: das Schaffen und Aufrechterhalten eines gemeinsamen Verständnisses, das angemessene Handeln, um das Problem zu lösen, und den Aufbau und das Aufrechterhalten der Teamorganisation. Das Programm ATC21S identifiziert ähnliche Dimensionen der Zusammenarbeit: Partizipation, Perspektivenübernahme und Regulation der Gefühle anderer.

Es gibt jedoch einen entscheidenden Unterschied zwischen diesen beiden Erfahrungen: Bei PISA interagierten die Schüler/innen mit Computeragenten, während ATC21S sich für einen Human-to-Human-Ansatz entschied. Die Wahl von PISA war durch das Ziel gerechtfertigt, die Messung zu standardisieren, um die Verwendung etablierter Messmethoden zu ermöglichen. Die Interaktion zwischen den Schüler/innen und dem Computeragenten beschränkte sich auf vordefinierte Aussagen im Multiple-Choice-Format, und jede mögliche Intervention der Schüler/innen war an eine bestimmte Antwort des Computeragenten oder ein Ereignis im Problemszenario gebunden. Diese stark kontrollierte Testumgebung und das Fehlen offener Antwortformate verringerten unweigerlich die Authentizität der Leistungsmessung.

Im Gegensatz dazu hat der Mensch-zu-Mensch-Ansatz von ATC21S mehr Aussagekraft, da die Schüler/innen selbst entscheiden konnten, wann und wie sie mit ihren Schulkolleg/innen über einen Chatbot interagierten. In dieser offeneren Umgebung ist es jedoch schwierig, das Verhalten der Schüler/innen vorherzusehen, was die Auswertung erschwert. Außerdem hängt der Erfolg eines Einzelnen vom Verhalten der anderen sowie deren Stimuli und Reaktionen ab. Dies führt zu dem Messproblem, wie man getrennte Scores für die einzelnen Schüler/innen und ihre Gruppe erstellen kann, und wirft die Frage auf, ob es fair ist, jemanden für die mangelnden Fähigkeiten oder die fehlende Motivation eines anderen zu bestrafen.

Diese Erfahrungen legen nahe, dass es möglich ist, sich eine nicht allzu ferne Zukunft vorzustellen, in der kollaborative Aufgaben ein integraler Bestandteil von Leistungsmessungen sind. Hu, Shubeck und Sabatini (2023, Kapitel 10 in *Innovating Assessments*) geben Beispiele dafür, wie die linguistische Datenverarbeitung (LDV) genutzt werden kann, um die Authentizität der Interaktion mit virtuellen Agenten zu erhöhen, indem intelligente Agenten entwickelt werden, die „verstehen", was Schüler/innen schreiben oder sagen, und entsprechend reagieren. In ähnlicher Weise haben Fortschritte in der LDV das Potenzial, die automatisierte Replikation von Experteneinschätzungen auf große Gesprächsdatensätze zu ermöglichen, wodurch die Qualität der aufgezeichneten Gespräche und schriftlichen Chats zwischen den Schüler/innen verbessert und die Analysekosten gesenkt werden. Unabhängig vom Ansatz erfordert die Realisierung authentischer kollaborativer Aufgaben parallel eine erhebliche Innovation bei der Messung, da Standardanalysemodelle nicht mit den zahlreichen zeit- und akteursübergreifenden Abhängigkeiten umgehen können, die in kollaborativen Umgebungen entstehen.

Quelle: Piacentini und Foster (2023) sowie Hu, Shubeck und Sabatini (2023), Kapitel 3 und 10 in *Innovating Assessments*.

SCHAFFUNG SOLIDER KONZEPTIONELLER GRUNDLAGEN

Mit größerer Klarheit über die Zielaktivitäten, -kontexte und -akteure für eine neue Art der Leistungsmessung ist es dann notwendig, eine Bestandsaufnahme jener Konzepte, Sprache und Werkzeuge zu anzufertigen, die die Menschen in der Zieldomäne verwenden, und die Merkmale einer guten Leistung in diesen Bereichen zu definieren. Bei der traditionellen Bewertung von Fächern (z. B. Mathematik) liegen bereits detaillierte Beschreibungen des Fachgebiets vor, die bei der Entwicklung der Leistungsmessung verwendet werden können. Wenn beispielsweise die Lesefähigkeit beurteilt werden soll, kann auf eine umfangreiche Literatur zurückgegriffen werden, in der die erforderlichen Kenntnisse und Fähigkeiten definiert sind und in der untersucht wurde, wie Kinder lesen lernen und dabei Fortschritte machen. In Bezug auf komplexe Kompetenzen wie kollaboratives Problemlösen oder Kommunikation ist jedoch nicht dasselbe Verständnis oder Wissen über Lernfortschritte vorhanden.

Um solche Informationen zu generieren, kann der Beitrag einer Gruppe von Experten hilfreich sein, welche in der Lage sind, neue Darstellungen dessen zu erstellen, was Expertise in diesen Bereichen bedeutet, und dabei so weit wie möglich auf empirische Beobachtungen zurückgreifen. Bei der kognitiven Aufgabenanalyse wird eine Vielzahl von Befragungs- und Beobachtungsstrategien, einschließlich der Prozessverfolgung, eingesetzt, um zu erfassen und zu beschreiben, wie Expert/innen komplexe Aufgaben ausführen (Clark et al., 2008). Eine etablierte Strategie der kognitiven Aufgabenanalyse ist beispielsweise die Methode der kritischen Ereignisse, bei der ein/e Expert/in gebeten wird, sich an die Entscheidungen zu erinnern, die er/sie in einer authentischen Situation getroffen hat, und diese zu beschreiben (siehe Kapitel 4 in *Innovating Assessments* für ein Beispiel für diese Praxis). Die erstellten Beschreibungen werden dann für die Entwicklung von Trainingserfahrungen und Messungen verwendet, da sie es ermöglichen, Merkmale von Aufgaben zu identifizieren, die geeignet sind, in die Identifikation von Entscheidungen, welche die meisten Hinweise auf die Kompetenz liefern, einbezogen zu werden.

Die Definition eines empirisch basierten Modells der Domäne kann durch Beobachtungsstudien darüber unterstützt werden, wie Schüler/innen an Aufgaben arbeiten, die die Zielfähigkeiten betreffen. Bei der Messung von Kollaborationsfähigkeiten können beispielsweise einige Modellaktivitäten für die Zusammenarbeit erstellt werden, die ein anfängliches Verständnis relevanter Situationen in dem Bereich widerspiegeln. Danach können Methoden der kognitiven Aufgabenanalyse zur Anwendung kommen, um diejenigen Schüler/innen zu identifizieren, die mehr oder weniger erfolgreich die Zusammenarbeit in Richtung des erwarteten Ergebnisses vorantreiben, und eine Bestandsaufnahme der Äußerungen und Handlungen von Schüler/innen auf verschiedenen Leistungsniveaus vorgenommen werden (z. B. wie sie Informationen innerhalb einer Gruppe austauschen, wie sie die Aufgabenteilung aushandeln usw.). Beobachtungsstudien schaffen Klarheit über die Abfolge von Handlungen, die durchgeführt werden müssen, um ein Leistungsziel zu erreichen, und liefern Beispiele für echte Arbeitsprodukte oder andere greifbare performanzbezogene Evidenz, die mit Aussagen über das Kompetenzniveau in Verbindung gebracht werden können.

In den nachfolgenden Entwicklungsphasen arbeiten die Entwickler von Leistungsmessungen mit Domänenexpert/innen zusammen, um die in ihrer Domänenanalyse gesammelten Informationen zu sog. *assessment arguments* zusammenzuführen; damit werden jene Aussagen bezeichnet, die sie über die Leistung der Schüler/innen treffen wollen,

die Daten, die als Beleg für diese Aussagen dienen sollen, und die Begründungen, die erklären, warum bestimmte Daten als angemessene Belege für eine bestimmte Aussage angesehen werden sollten (Toulmin, 1958; Mislevy und Riconscente, 2006). Wie in Kasten 4 veranschaulicht, können Argumente mit Hilfe von „Entwurfsmustern" formalisiert werden, die das Wissen, die Fähigkeiten und die Einstellungen der Lernenden beschreiben, die im Mittelpunkt der Leistungsmessung stehen, sowie die potenziellen Beobachtungen, Arbeitsprodukte und Bewertungsraster, die Testentwickler verwenden möchten, und die Merkmale potenzieller Testaufgaben. Diese Entwurfsmusterstruktur hilft bei der Identifikation und Konsolidierung der konzeptionellen Grundlagen einer Leistungsmessung und dient als Ausgangsbasis für die Ausarbeitung der technischen Spezifikationen, die die Operationalisierung der Leistungsmessung leiten – d. h. die Schüler-, Aufgaben- und Evidenzmodelle des ECD-Rahmens in Abbildung 3.

KASTEN 4.

ENTWURFSMUSTER IN DER LEISTUNGSMESSUNG: EIN BEISPIEL DER PISA-STUDIE

Entwurfsmuster für die computergestützte Modellierung im Rahmen der PISA-Erhebung 2025 Lernen in der digitalen Welt

Begründung	Das Modellieren ist eine der wichtigsten Praktiken des wissenschaftlichen Denkens, aber Schüler/innen beschäftigen sich während der Pflichtschulzeit nur selten damit. Computer machen das Modellieren für Lernende, insbesondere für Anfänger, leichter zugänglich und sinnvoller. Die Beobachtung, wie Schüler/innen Computermodelle erstellen, verfeinern und verwenden, liefert relevante und interpretierbare Hinweise darauf, wie fähig Schüler/innen sind, ihr eigenes Wissen und Verständnis komplexer Phänomene mithilfe von Computern zu entwickeln.
Schwerpunktkenntnisse, -fähigkeiten und -einstellungen	• Verständnis des Konzepts der Variablen, einschließlich abhängiger, unabhängiger, Kontroll- und Moderatorvariablen • Erstellung einer abstrakten Darstellung eines Systems, das von einem Computer ausgeführt werden kann; Sicherstellung, dass das Modell wie erwartet funktioniert (z. B. Beobachtung des Verhaltens von Agenten in einer auf dem Modell basierenden Simulation) • Erkennen von Trends, Anomalien oder Korrelationen in Daten • Experimentieren mit der Variablenkontrollstrategie • Verwendung eines Berechnungsmodells, um Vorhersagen über das Verhalten eines Systems zu treffen
Zusätzliche Kenntnisse, Fähigkeiten und Einstellungen	• Funktionale Kenntnisse der Informations- und Kommunikationstechnik (IKT) • IKT-Selbstwirksamkeit • Vorwissen über das zu modellierende Phänomen • Ausdauer, Gewissenhaftigkeit und Mastery-Orientierung
Mögliche Beobachtungen und Arbeitsprodukte	• Das Schülermodell stellt die verfügbaren Informationen über die reale Situation dar. • Die Schüler/innen konsultieren relevante Informationsquellen und sammeln relevante Daten, um die Modellparameter festzulegen. • Die Schüler/innen modifizieren ein unvollständiges oder fehlerhaftes Modell und begründen ihre Änderungen. • Die Schüler/innen identifizieren die Schwächen des Modells. • Die Schüler/innen verwenden ihr Modell, um korrekte Vorhersagen zu treffen (unter Berücksichtigung der verfügbaren Daten).
Charakteristische Merkmale der Aufgaben	• Die Schüler/innen erhalten entweder Informationen über ein reales soziales oder wissenschaftliches Phänomen, das sie modellieren sollen, oder sie erhalten die Werkzeuge, um diese Informationen zu beschaffen. • Die Schüler/innen können ihr Modell überprüfen, indem sie dessen Ergebnisse mit realen Daten vergleichen. • Die Schüler/innen können das Modell verwenden, um Vorhersagen zu treffen.
Variable Merkmale der Aufgaben	• Grad der Vertrautheit mit dem zu modellierenden Phänomen • Komplexität der für die Modellierung verwendeten IKT-Tools • Die Schüler/innen verbessern ein (ihnen zur Verfügung gestelltes) Grundmodell oder bauen das Modell von Grund auf. • Die Schüler/innen müssen relevante Daten finden (z. B. in einer Informationsquelle) oder ihre eigenen Daten durch Experimentieren erstellen. • Anzahl der zu modellierenden Variablen und Struktur des Systems (einfach vs. mehrstufig)
Sachzwänge und Herausforderungen	• Begrenzte Zeit zum Erlernen der Verwendung des Modellierungswerkzeugs • Begrenzte Zeit zum Erlernen ungewohnter Modellierungskonzepte (z. B. Variablenkontrollstrategie) • Große Unterschiede im Vorwissen der Schülerpopulation, d. h. es ist schwierig, alle Schüler/innen bei derselben Aufgabe angemessen zu fordern.

Quelle: Piacentini (2023), Kapitel 6 in *Innovating Assessments*.

BERÜCKSICHTIGUNG SOZIOKULTURELLER UNTERSCHIEDE BEI DER DEFINITION VON BEWERTUNGSKONSTRUKTEN

Um vergleichende Schlüsse ziehen zu können, sind die Äquivalenz der Messung und die Vergleichbarkeit der Ergebnisse unbedingt erforderlich, wenn Tests in mehreren Sprachen durchgeführt werden oder wenn Schüler/innen aus verschiedenen kulturellen Gruppen Tests in derselben Sprache ablegen. Fragen der kulturübergreifenden Validität und Vergleichbarkeit sind von besonderer Bedeutung für die Messung komplexer Konstrukte in multikulturellen und mehrsprachigen Kontexten, wie z. B. bei internationalen Leistungsmessungen und Leistungsmessungen in Ländern mit verschiedenen kulturellen Bevölkerungsgruppen.

Die Äquivalenz der Konstrukte ist ein wichtiger Aspekt, der bei der Festlegung des Schwerpunkts einer Leistungsmessung beachtet werden muss. Sie ist das Ausmaß, in dem die Definitionen der Konstrukte für die Zielgruppen der Leistungsmessung ähnlich sind, ob von den Schüler/innen erwartet wird, dass sie diese Konstrukte auf ähnliche Weise entwickeln und Fortschritte machen, und ob die Konstrukte für alle Bevölkerungsgruppen auf ähnliche Weise zugänglich sind. Sie ist für alle Leistungsmessungen, die für multikulturelle und mehrsprachige Gruppen bestimmt sind, von entscheidender Bedeutung, besonders relevant ist sie aber bei groß angelegten Messungen komplexer und mehrdimensionaler Konstrukte (Ercikan und Oliveri, 2016).

Konstrukte wie Kreativität, Intelligenz, kritisches Denken und Zusammenarbeit werden in den Schulen nicht einheitlich gelehrt und in verschiedenen Kulturen unterschiedlich konzeptualisiert und definiert. So unterscheiden sich beispielsweise die Entwicklung der Kreativität und die Ausprägung kreativer Verhaltensweisen in den verschiedenen kulturellen Gruppen (Lubart, 1990; Niu und Sternberg, 2001). Laut anderen Forschenden sind die Konzepte der Intelligenz in kulturellen Kontexten begründet und werden die Konstrukte als solche in diesen Kontexten unterschiedlich definiert (Sternberg, 2013).

Da komplexe Fähigkeiten in soziale Kontexte eingebettet und von kulturellen Normen und Erwartungen geprägt sind, ist zu erwarten, dass ihre Ausprägung und der Wert, der den Leistungen der Schüler/innen zugeschrieben wird, von Kultur zu Kultur variieren. Aufgrund dieser Unterschiede zwischen den kulturellen Gruppen ist klar zu bestimmen, welche Aspekte eines Konstrukts in einem vergleichenden Kontext sinnvoll gemessen und somit in die Leistungsmessung einbezogen werden können, auch wenn dies zu einer gewissen Verengung des Konstrukts führen könnte. Die PISA-Erhebung 2022 des kreativen Denkens (OECD, 2022) ist ein Beispiel dafür, wie sich die Messung eines komplexen Konstrukts über Sprach- und Kulturgruppen hinweg dennoch auf bestimmte Aspekte des Konstrukts konzentrieren kann, die die Vergleichbarkeit optimieren (siehe Kasten 5).

KASTEN 5.

BERÜCKSICHTIGUNG SOZIOKULTURELLER UNTERSCHIEDE BEI DER KONSTRUKTDEFINITION VON GROSS ANGELEGTEN LEISTUNGSMESSUNGEN

Sicherstellung der Konstruktäquivalenz in der PISA-Erhebung 2022 des kreativen Denkens

In der PISA-Domäne des kreativen Denkens (2022) wird betont, dass sich die Messitems auf Wissen und Erfahrungen stützen sollten, die allen Schüler/innen auf der ganzen Welt gemeinsam sind und für welche Schüler/innen innerhalb der Einschränkungen einer PISA-Erhebung in sinnvoller und realistischer Weise kreative Arbeiten produzieren können.

Um dies zu gewährleisten, wurden insbesondere fünf Punkte berücksichtigt:

- Die Messung konzentrierte sich auf das engere Konstrukt des kreativen Denkens (statt auf das umfassendere Konstrukt der Kreativität), definiert als die Kompetenz, vielfältige, kreative Ideen zu produzieren, zu evaluieren und zu verbessern. Dieser engere Fokus betonte die kognitiven Prozesse im Zusammenhang mit der Ideengenerierung, während Kreativität auch Persönlichkeitsmerkmale umfasst und subjektive Beurteilungen des kreativen Werts der Antworten der Schüler/innen erfordert.

- Es wurde eine Definition kreativen Denkens erstellt sowie definiert, wodurch es ermöglicht wird (d. h. Indikatoren für Möglichkeiten, kreatives Denken zu erlernen) und wie es im Kontext von 15-Jährigen im Klassenzimmer aussieht, wobei der Schwerpunkt auf jenen Aspekten des Konstrukts liegt, die eher in schulischen Kontexten (als außerhalb der Schule) entwickelt werden können.

- Es wurden kulturübergreifend relevante Domänen identifiziert, mit denen sich 15-Jährige beschäftigen und in welchen sie mit hoher Wahrscheinlichkeit kreatives Denken geübt haben (z. B. Schreiben von Kurzgeschichten, Erstellen visueller Produkte, Brainstorming zu allgemeinen sozialen und wissenschaftlichen Problemen).

- Der Schwerpunkt der Messung lag auf der Originalität der Ideen (definiert als statistische Seltenheit) und auf ihrer Vielfalt (definiert als Zugehörigkeit zu verschiedenen Ideenkategorien) und nicht auf ihrem kreativen Wert (der eher von soziokulturellen Unterschieden abhängt).

- Die Bewertungsraster, anhand derer die Antworten durch menschliche Bewerter ausgewertet wurden, wurden in erheblichem Umfang kulturübergreifend überprüft und mit einer Analyse von Antwortbeispielen von Schüler/innen aus mehreren Ländern verfeinert.

Quelle: OECD (2022), Thinking Outside the Box: The PISA 2022 Creative Thinking Assessment, https://www.oecd.org/pisa/innovation/creative-thinking/.

ERNEUERUNG DER *BEOBACHTUNGS*KOMPONENTE: EINBEZUG VIELFÄLTIGERER UND INTERAKTIVER AUFGABEN

Wenn Leistungsmessung als ein Prozess betrachtet wird, bei dem man aus Belegen schlussfolgert, müssen die gestellten Aufgaben relevante Belege aus den Schüler/innen hervorbringen, und diese Belege müssen eindeutig mit dem Konstrukt verbunden sein. Mit anderen Worten, die Testaufgaben oder -situationen sollten die Beobachtung jener Leistungstypen ermöglichen, die von den Schüler/innen erwartet werden. Bei Konstrukten wie dem mathematischen Wissen ist die Verbindung zwischen den Testindikatoren und dem Konstrukt ziemlich direkt: Eine korrekte Antwort auf eine bestimmte Frage zeigt, dass die Schüler/innen das Thema beherrschen. Diese Logik reicht jedoch möglicherweise nicht aus, um die Komplexität der Kompetenzen des 21. Jahrhunderts zu erfassen.

Ein zentrales Argument von *Innovating Assessments* ist, dass Leistungsmessungen eher valide Evidenz dafür liefern, was Schüler/innen wissen und können, wenn sie diese mit authentischen Situationen konfrontieren. Grund für den Ruf nach Innovation ist die Tatsache, dass die bestehenden Tests dies oft nicht ermöglichen – zum Teil, weil sich die technischen Möglichkeiten, eine solche Vision in großem Maßstab umzusetzen, nur langsam entwickelt haben. Bildungsbewertungen, insbesondere groß angelegte standardisierte Tests, wurden innerhalb einer Reihe von Beschränkungen entwickelt – Druck- und Transportkosten, Testsicherheit, Testumgebung, Testzeit und Auswertungskosten – bei gleichzeitiger Einhaltung psychometrischer Standards für Zuverlässigkeit, Gültigkeit, Vergleichbarkeit und Fairness. Die Hauptmerkmale der „traditionellen" Testentwicklung, -verwaltung, -auswertung und -dokumentation, wie z. B. Multiple-Choice-Aufgaben, haben sich aufgrund solcher Einschränkungen herausgebildet (OECD, 2013), und ihre Fähigkeit, komplexere und facettenreichere Leistungsaspekte zu erfassen, ist dementsprechend begrenzt geblieben.

Dennoch gelten viele der Einschränkungen bei der Testentwicklung und -durchführung entweder nicht mehr, haben sich geändert oder können zum großen Teil aufgrund der technologischen und datenanalytischen Fortschritte gelockert werden. Insbesondere der digitale Werkzeugkasten, der zur Verfügung steht, erweitert die Möglichkeiten der Testentwicklung dramatisch und birgt das Potenzial, Testerfahrungen weniger künstlich und valider zu gestalten, indem die Situationen oder Kontexte, in denen die Zielkonstrukte im wirklichen Leben verwendet werden, angenähert oder simuliert werden.

AUFGABENDESIGN NEU ÜBERDENKEN

Piacentini, Foster und Nunes (2023) geben eine Reihe von Empfehlungen für das Design von Aufgaben und Items (Kapitel 2 in *Innovating Assessments*). Dazu gehören (1) erweiterte Performanzaufgaben mit niedrigen Untergrenzen („low floors") und hohen Obergrenzen („high ceilings"); (2) die explizite Berücksichtigung von Fachwissen; und (3) die Bereitstellung von Gelegenheiten für produktives Scheitern und Lernen im Test sowie das Angebot von Feedback und didaktischer Unterstützung während des Tests. Die Idee hinter diesen Grundsätzen ist

nicht, traditionellere Formen von Bewertungserfahrungen und Antwortformaten abzuschaffen, da diese immer noch relevante Informationen für einige interpretative Zwecke liefern können (z. B. zur Identifikation von Wissenslücken). Vielmehr geht es darum, diese etablierten Formen der Leistungsmessung durch eine andere Art von Testerfahrungen zu ergänzen, die diese innovativen Merkmale einbeziehen.

Designprinzip I: Erweiterte, performanzbezogene „Low-Floor-High-Ceiling"-Aufgaben

Bei der Leistungsmessung, insbesondere bei groß angelegten summativen Tests, haben Effizienzüberlegungen dazu geführt, dass kurzen, diskreten Aufgaben gegenüber längeren Performanzaktivitäten der Vorzug gegeben wird. Im Allgemeinen liefert die Verwendung vieler kurzer Aufgaben zuverlässigere Daten darüber, ob die Schüler/innen bestimmte Kenntnisse beherrschen und eine Reihe vorgegebener Vorgangsweisen befolgen können, da die Informationen über eine größere Anzahl von Beobachtungen akkumuliert werden. Auch die Messung ist einfacher: Die Erkenntnisse werden durch die Anwendung etablierter psychometrischer Modelle auf völlig unabhängige Items gewonnen. Wenn sich jedoch der Zweck der Leistungsmessung dahingehend verschiebt, dass gemessen wird, ob die Schüler/innen neues Wissen in einem Umfeld mit vielen Wahlmöglichkeiten aufbauen können, dann sollten den Schüler/innen Testaufgaben und -umgebungen vorgelegt werden, die für dieses Ziel geeignet sind.

Dazu ist es wichtig zu überlegen, wie ein Test den Schüler/innen eine Herausforderung bieten kann, die zielgerichtet ist und ihnen genügend Zeit lässt, ihre Kompetenzen zu demonstrieren. Die Einbeziehung erweiterter Einheiten, bei denen mehrere Aktivitäten als Schritte zur Erreichung eines Hauptziels aneinandergereiht werden, kann den Schüler/innen eine authentischere und motivierendere Testerfahrung bieten. Werden die Testteilnehmenden dazu ermutigt, ihre Denkweise von „Ich muss so viele Aufgaben wie möglich richtig beantworten" zu „Ich habe eine Herausforderung, auf die ich hinarbeiten und die ich bewältigen muss" zu ändern, könnte dies letztlich einen valideren Beleg dafür liefern, wozu die Schüler/innen außerhalb der Zwänge stressiger und zeitkritischer Testkontexte in der Lage sind.

Ausgedehnte, performanzbasierte Aufgaben sind schwieriger zu konzipieren, nicht zuletzt, weil man eine kohärente Handlung aufbauen muss, die die Schüler/innen bei der Stange hält, und weil man potenzielle Abhängigkeitsprobleme angehen muss – zum Beispiel, indem man Rettungsanker anbietet, um Schüler/innen, die Probleme haben, von einer Aktivität zur nächsten zu bringen. Gleichzeitig sollten die Tests es allen Schüler/innen ermöglichen, ihre Lernfähigkeit und ihren Fortschritt zu demonstrieren, unabhängig von ihrem anfänglichen Wissens- oder Fähigkeitsniveau, indem sie Aufgaben konzipieren, die eine niedrige Untergrenze („low floor") und eine hohe Obergrenze („high ceiling") haben, d. h., dass sie für alle Schüler/innen zugänglich sind und dennoch die besten Leistungen fordern (siehe Kasten 6 für ein Beispiel aus der PILA-Plattform der OECD).

Eine Möglichkeit, solche Low-Floor-High-Ceiling-Probleme zu entwickeln, besteht darin, die Schüler/innen zu bitten, ein originelles Artefakt zu schaffen: Das kann eine Geschichte, ein Spiel, ein Design für ein neues Produkt sein, ein Untersuchungsbericht über eine Nachricht, eine Rede usw. Diese offeneren Performanzaufgaben erzeugen ein breites Spektrum an qualitativ unterschiedlichen Antworten, und selbst Spitzenkräfte

haben Anreize, Ressourcen zu nutzen, die ihnen helfen können, eine reichhaltigere, vollständigere und einzigartige Lösung zu finden. Das Konzept „Low Floor, High Ceiling" kann auch im Zusammenhang mit standardisierten Problemlösungsaufgaben eingesetzt werden, um den Schüler/innen zu verdeutlichen, dass es Zwischenziele gibt und dass von ihnen erwartet wird, dass sie so weit wie möglich auf eine anspruchsvolle Lösung hinarbeiten.

Adaptive Designs können auch die Komplexität der Messung des Lernens in Aktion bei heterogenen Schülerpopulationen berücksichtigen. Eine relativ einfache Möglichkeit besteht darin, Szenarien zu schaffen, in denen die Lernenden ein komplexes Ziel zu erreichen haben und auf dem Weg dorthin eine Reihe von Aufgaben lösen, deren Schwierigkeitsgrad allmählich ansteigt (ähnlich wie beim „Levelling-up" in Videospielen). Leistungsstärkere Schüler/innen werden die anfänglichen einfachen Aufgaben schnell lösen und danach auf Probleme stoßen, die sie herausfordern; weniger gut vorbereitete Schüler/innen werden sich mit den einfacheren Aufgaben beschäftigen können, auch wenn sie nicht die gesamte Sequenz lösen. Bei einem solchen Design arbeiten beide Gruppen von Schüler/innen an der Spitze ihrer Fähigkeiten, was offensichtliche Vorteile in Bezug auf die Messqualität und die Testbeteiligung mit sich bringt. Mit den heutigen Technologien könnte dieses Design noch weiter verbessert werden, indem mehrere adaptive Pfade innerhalb eines Szenarios eingeführt werden: Je nach Qualität ihrer Arbeit werden die Schüler/innen on-the-fly zu leichteren oder schwierigeren Teilaufgaben geleitet.

KASTEN 6.

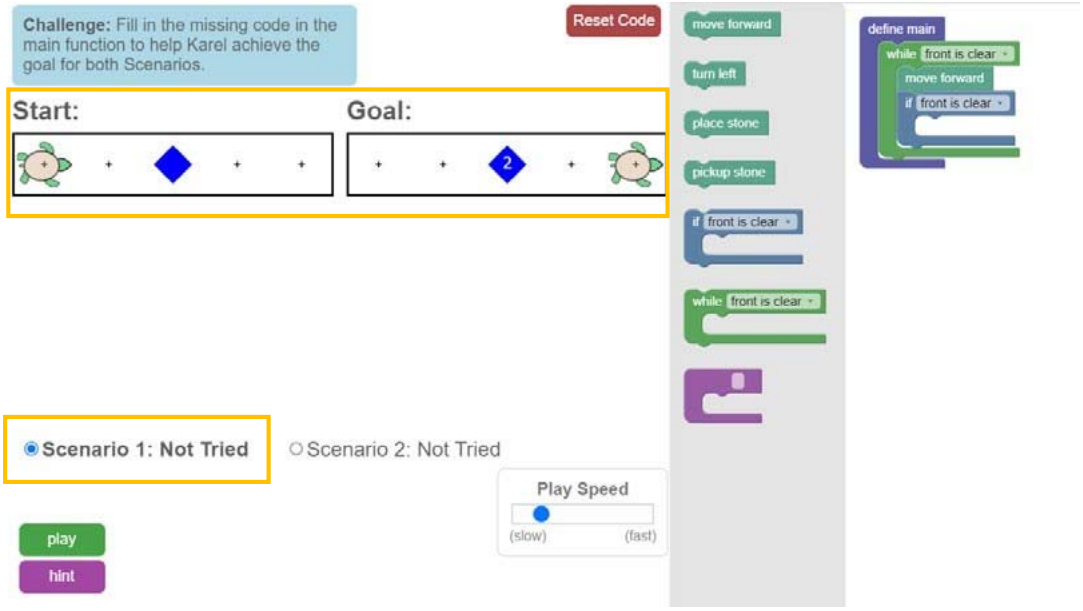
TESTAUFGABEN MIT „NIEDRIGER UNTERGRENZE" („LOW FLOOR") UND „HOHER OBERGRENZE" („HIGH CEILING")

Berücksichtigung von Schüler/innen mit besonderen Fähigkeiten bei der PILA-Bewertung des rechnerischen Problemlösens

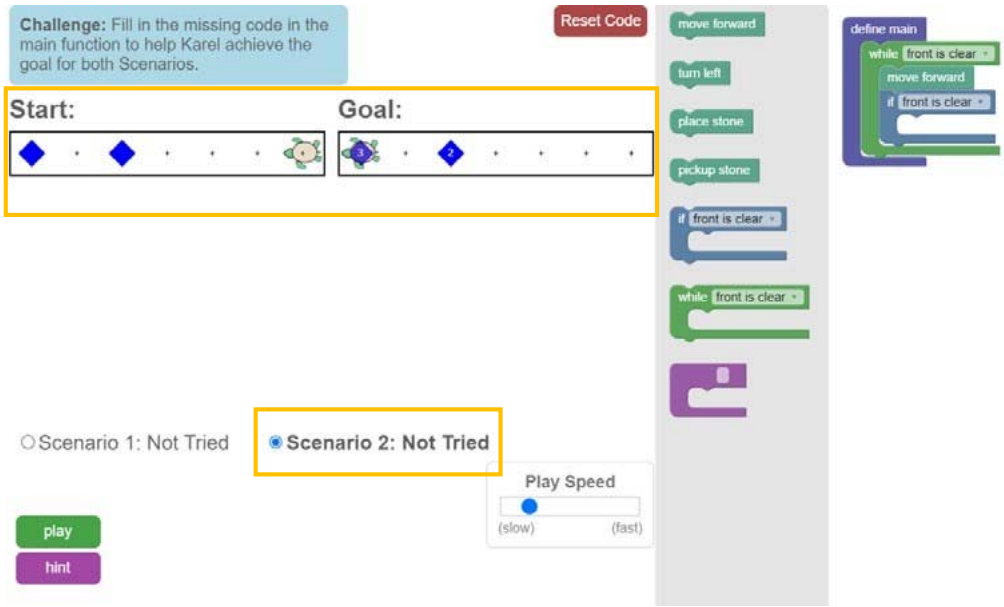
Die Platform for Innovative Learning Assessments (PILA) ist ein von der OECD koordiniertes Forschungslabor. Die Bewertungen in PILA sind als Lernerfahrungen konzipiert und bieten Echtzeit-Feedback über die Fortschritte der Schüler. Sie können also auch im Rahmen des Unterrichts eingesetzt werden. Ein übergeordnetes Ziel von PILA ist es, Entwickler von Leistungsmessungen, Programmierer, Messexperten und Pädagogen zusammenarbeiten zu lassen, um neue Wege zu finden, die Lücke zwischen Lernen und Leistungsmessung zu schließen.

Eine in PILA entwickelte Anwendung konzentriert sich auf die rechnerische Problemlösung. Die Schüler/innen verwenden eine blockbasierte visuelle Programmierschnittstelle, um einen Schildkrötenroboter („Karel") anzuweisen, bestimmte Aktionen auszuführen. Dieser Test hat eine niedrige Untergrenze und eine hohe Obergrenze: Die Intuitivität der visuellen Sprache und die eingebetteten Lehrmittel (z. B. interaktives Tutorial, Arbeitsbeispiele) ermöglichen es Schüler/innen, die keine Programmiererfahrung haben, sich erfolgreich mit einfachen algorithmischen Aufgaben zu beschäftigen. Die gleiche Umgebung kann jedoch auch dazu verwendet werden, Probleme zu schaffen, die selbst für erfahrene Programmierer eine Herausforderung darstellen.

Die Bilder unten zeigen ein Beispielproblem, bei dem die Schüler/innen ein einziges Programm erstellen sollen, mit dem Karel in zwei verschiedenen Szenarien das Ziel erreicht. Um das Problem zu lösen, können die Schüler/innen zwischen den beiden Szenarien hin- und herschalten, um die Unterschiede in der Umgebung visuell zu beobachten, sowie das Ausmaß, in dem ihr Programm das Problem in beiden Szenarien löst. Bei dieser Art von Aufgaben entwickeln selbst Schüler/innen mit soliden Programmierkenntnissen in der Regel mehrere Iterationen ihres Programms, bevor sie eine Lösung finden. Die Bewertungsmodelle berücksichtigen Teillösungen (z. B. die Fähigkeit eines Schülers, das Problem in einer Welt zu lösen), und die Berichts-Dashboards enthalten komplexere Leistungsindikatoren (z. B. die Anzahl der Iterationen, die die Schüler/innen getestet haben).

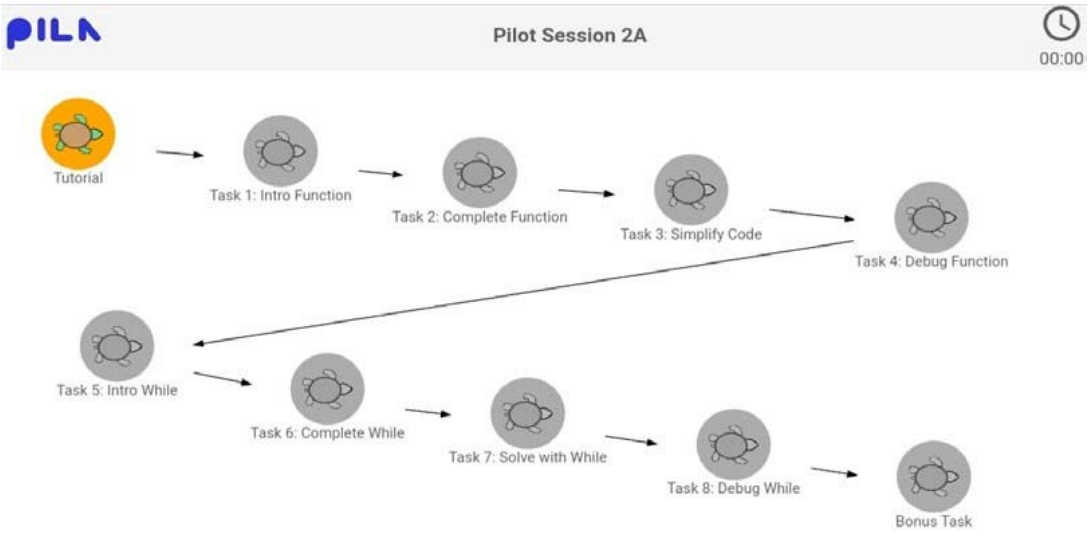


Aufgabe: Die Schüler müssen Karel so programmieren, dass er sich vorwärtsbewegt und einen Stein auf dem Weg ablegt, so dass er dem Zielzustand von Szenario 1 entspricht (Bild oben). Derselbe Code soll auch Szenario 2 lösen (Bild unten).



Jeder PILA-Test ist außerdem als eine Abfolge zunehmend komplexer Aufgaben strukturiert, die ein gemeinsames Lernziel haben (z. B. die effiziente Nutzung von Funktionen). Entwickler/innen von Leistungsmessungen und Lehrkräfte haben die Möglichkeit, die Schüler/innen so lange an eine bestimmte Aufgabe zu binden, bis sie in der Lage sind, diese zu lösen (= „Level-up"-Mechanismus), oder die Schüler/innen können selbst bestimmen, wie sie sich in der Aufgabensequenz bewegen. Nur von hoch qualifizierten Schüler/innen wird erwartet, dass sie die gesamte Aufgabensequenz abschließen, und dies wird den Schüler/innen zu Beginn klar mitgeteilt, um Frustrationserlebnisse zu vermeiden. Für die Zukunft plant PILA die Einführung von adaptiven Pfaden (d.h. Problemsequenzen, die sich in Echtzeit an die Leistung der Schüler/innen anpassen), um die Erfahrung noch stärker an das Vorwissen und die Fähigkeiten der Schüler/innen anzupassen.

Die folgende Abbildung zeigt ein Beispiel für eine Testabfolge („Map") in der Anwendung „Karel".



Quelle: Piacentini, Foster und Nunes (2023), Kapitel 2 in *Innovating Assessments*.

Designprinzip II: Explizite Berücksichtigung von Fachwissen

Wie bereits erörtert, ist es bei der Entwicklung von Messungen der Kompetenzen des 21. Jahrhunderts wichtig, das Wissen, das die Schüler/innen für eine sinnvolle Beschäftigung mit den Testaktivitäten benötigen, explizit zu identifizieren und einzuschätzen, inwieweit Unterschiede im Vorwissen die Evidenz bezüglich der Zielfähigkeiten beeinflussen. Im Zusammenhang mit groß angelegten, summativen Leistungsmessungen könnten allgemeine Behauptungen wie „Schüler in Land A lösen Probleme besser als Schüler in Land B" irreführend sein. Tatsächlich kann bei einer einzigen summativen Leistungsmessung nur behauptet werden, dass die Schüler in Land A besser als die Schüler in Land B Probleme in den im Test dargestellten Situationen lösen können (höchstwahrscheinlich einer begrenzten Anzahl von Situationen, die in einem oder wenigen Wissensbereichen kontextualisiert sind).

Die Messung des relevanten Wissens, über das die Schüler/innen verfügen, wenn sie eine Performanzaufgabe lösen (z. B. mittels einer kurzen Reihe von Aufgaben zu Beginn des Tests), sollte ein integraler Bestandteil des Entwurfs- und Evaluierungsprozesses von Leistungsmessungen der nächsten Generation werden. Diese Informationen können auch dazu beitragen, das Verhalten und die Entscheidungen der Schüler/innen bei Tests mit komplexen Performanzaufgaben zu interpretieren. Es könnte auch versucht werden, die Variabilität des relevanten Vorwissens zu minimieren, indem den Schüler/innen Anleitungen, Beispiele und Übungsaufgaben zur Verfügung gestellt werden, die ihnen bei der Beschäftigung mit einer Aufgabe helfen. Diese Ansätze können sowohl für die Berücksichtigung des Fachwissens als auch für das Wissen über die in die Testumgebung eingebetteten Ressourcen oder Tools nützlich sein (d. h. sie helfen den Schüler/innen, sich in der Testumgebung zurechtzufinden).

Designprinzip III: Gelegenheiten für produktives Scheitern und Lernen im Test, Feedback und Unterstützungsmechanismen

Bei herkömmlichen Tests besteht das Ziel darin, das vor der Aufgabe erworbene Wissen der Schüler/innen zu messen. In der Regel wird den Schüler/innen kein Feedback gegeben, die Aufgaben sind häufig sehr unterschiedlich (um zu vermeiden, dass die Antworten im selben Test verraten werden), und die Antworttypen beschränken sich meist auf kategorische Antworten, d. h. richtige oder falsche Antworten. Diese Instrumente sind unzureichend, wenn die Ziele von der Messung der Anwendung vorhandenen, statischen Wissens (*Lernergebnisse*) auf die Messung der Dynamik des Erwerbs und der Entwicklung neuen Wissens (*Lernprozesse*) bei der Bewältigung komplexer Aufgaben ausgeweitet werden.

Eine vielversprechende Methode zur Behebung der derzeitigen Mängel ist der Einsatz von „Erfinderaktivitäten" bei Leistungsmessungen, bei denen die Schüler/innen Probleme lösen sollen, die scheinbar nichts mit dem Unterrichtsstoff zu tun haben und Konzepte oder Vorgangsweisen beinhalten, die sie noch nicht gelernt haben. Die Schüler/innen müssen für diese neuartigen Probleme ihre eigenen originellen Lösungen erfinden, in diesem Prozess neigen sie dazu, Fehler zu machen, und finden meist keine kanonischen Lösungen. Erfinderaktivitäten helfen den Schüler/innen jedoch dabei, Konzepte in ihrer Tiefe zu verstehen, alte Interpretationen und Vorgangsweisen loszulassen, wenn sie nicht funktionieren, und nach neuen Mustern und Interpretationen zu

suchen – und können im Rahmen einer Leistungsmessung Aufschluss darüber geben, ob die Schüler/innen ihr Wissensschema flexibel auf unbekannte Kontexte anwenden können, so wie es lernfähige Expert/innen tun. Sicherlich ist eine völlig offene und ungelenkte Erkundung nicht unbedingt der beste Beleg dafür, was Anfänger tun können; die Lernaktivitäten müssen dennoch sorgfältig konzipiert werden, um die Schüler/innen beim Aufbau ihres Verständnisses zu unterstützen, während sie Probleme mit unbekannten Aspekten erfinden und mit ihnen interagieren.

Leistungsmessungen der nächsten Generation sollten Anleitungen und Hilfestellungen während des Lösungsprozesses in Form von Ratschlägen, Feedback oder Aufforderungen berücksichtigen. Solche Hilfestellungen können verschiedene Funktionen erfüllen: (1) Wecken des Interesses der Schüler/innen, wenn sie unmotiviert erscheinen; (2) Verbesserung des Verständnisses für die Anforderungen der Aufgabe, wenn die Schüler/innen Verwirrung zeigen; (3) Reduktion der Freiheitsgrade oder der Anzahl der für eine Lösung erforderlichen Teilhandlungen; (4) Beibehalten der Richtung; (5) Markierung kritischer Merkmale, einschließlich Abweichungen zwischen dem, was ein/e Schüler/in produziert hat, und dem, was er/sie als richtig anerkennen würde; (6) Demonstration oder Modellierung von Lösungen, z. B. Reproduktion und Vervollständigung einer Teillösung, die ausprobiert wurde; und (7) Anregung zu Artikulation und Reflexion (Guzdial, Rick und Kehoe, 2001).

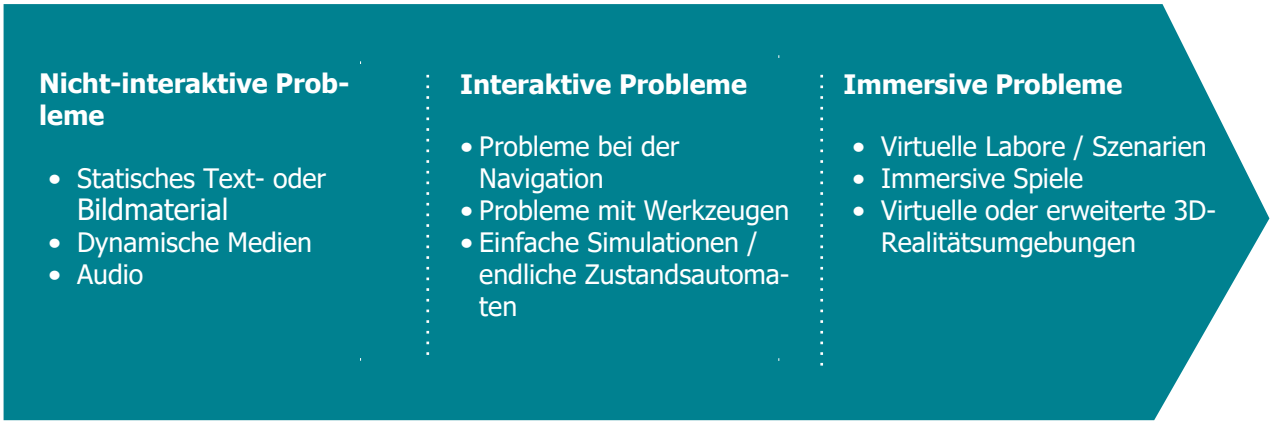
NUTZUNG MODERNER TECHNOLOGIEN FÜR EIN INNOVATIVES DESIGN VON LEISTUNGSMESSUNGEN

Die laufenden technologischen Entwicklungen machen die oben genannten Innovationen zunehmend möglich, indem sie die für das Testdesign verfügbaren Instrumente erweitern. Wie von Sabatini et al. erörtert (*Innovating Assessments*, Kapitel 7), erweitern moderne Technologien die Bandbreite des Möglichen, wenn es um das Design von Aufgabenformaten, Testmerkmalen und Evidenzquellen geht.

Aufgabenformat: Von statischen zu interaktiven und dynamischen Bewertungssituationen

Viele Leistungsmessungen sind durch nicht-interaktive Aufgaben gekennzeichnet. Diese enthalten oft statischen schriftlichen Text oder visuelle Stimuli (z. B. Fotos, Zeichnungen, Tabellen, Karten, Diagramme oder Tabellen) und in einigen Fällen dynamischere Stimuli wie Audio, Animationen, Videos und andere multimediale Inhalte. Bei nicht-interaktiven Aufgaben liefert das Stimulusmaterial den Schüler/innen in der Regel alle Informationen, die sie zur Lösung der Aufgabe benötigen, die Antworten erfolgen häufig in Form von schriftlichen oder eng begrenzten Aufgaben, bei denen wenig bis gar keine Interaktivität der Testteilnehmenden möglich ist, und die Testumgebung entwickelt sich nicht mit der Interaktion der einzelnen Testteilnehmenden weiter.

Abbildung 5. Kontinuum der Aufgabenformate
Nicht-interaktive, interaktive und immersive Bewertungsprobleme



Quelle: Sabatini et al. (2023), Kapitel 7 in *Innovating Assessments.*

Im Gegensatz dazu ermöglichen interaktive Aufgaben den Schüler/innen, sich aktiv an den Handlungsprozessen zu beteiligen, indem sie die für komplexere Arten von Leistungen typischen Problemlösungsszenarien schaffen. Diese Aufgabenformate sind offener und reagieren stärker auf die Handlungen und Verhaltensweisen der Testteilnehmenden. Sie sind in der Regel mehrstufig, beinhalten den Einsatz von Computeranwendungen, Werkzeugen oder Suchmaschinen, was den heutigen Praxiszusammenhang besser widerspiegelt, und erfordern in der Regel eine Navigation innerhalb von sowie zwischen Bildschirmanzeigen.

Mit Hilfe der Technologie können auch wirklich immersive Probleme in die Leistungsmessung einbezogen werden. Dazu gehören simulierte Labore, immersive Spiele oder 3D-Modellierungen und virtuelle Realitäten. Immersive Aufgaben ermöglichen es den Schüler/innen, durch eine zwei- oder dreidimensionale Darstellung einer virtuellen Welt – imaginär oder real – auf einem Bildschirm oder über Virtual-Reality-Headsets zu navigieren. Immersive Aufgaben verwenden häufig spielerische Elemente, um die Motivation zu steigern und die Erfahrung der Lernenden zu unterstützen oder zu kontrollieren (Pellas et al., 2018). Zu den Beispielen gehören Simulationen, die vor allem für Berufsausbildungen verwendet werden, wie z. B. virtuelle Flugsimulationen oder Simulationen medizinischer Eingriffe, obwohl diese Arten von Aufgaben zunehmend auch in großem Umfang konzipiert und umgesetzt werden können.

Wichtig ist, dass eine größere Interaktivität und Immersivität der Testaufgaben mit konstruktbezogenen und praktischen Überlegungen in Einklang gebracht werden muss. Aufgaben, die sich im Laufe der Interaktion mit den Testteilnehmenden weiterentwickeln, können zu weniger einheitlichen Aufgabenerfahrungen und damit zu einer ungleichmäßigen Erfassung der Zielkonstrukte führen, was Rückschlüsse auf verschiedene Schülerpopulationen erschwert. Authentische und interaktive Aufgaben können auch mehr Zeit in Anspruch nehmen als einfachere statische Aufgaben. Das Design von interaktiven und immersiven Aufgaben beinhaltet notwendigerweise die Optimierung der Balance zwischen der Authentizität der Aufgabe und den Einschränkungen: Bei immersiven Designs ist es von größter Bedeutung, dass die Aufgaben in der virtuellen Welt auf Leistungsunterschiede zwischen einzelnen Schüler/innen (z. B. Anfängern und Experten) reagieren, sodass sie das Wissen und die Fähigkeiten von Relevanz wirklich widerspiegeln (d. h. dass sie Konstruktvalidität besitzen) und dass sie die Schüler/innen nicht von der eigentlichen Aufgabe ablenken.

Testmerkmale: Einführung von Testadaptivität und Lernressourcen

Die digitale Technologie kann auch zur Innovation von Testmerkmalen dienen, die sich auf die Möglichkeiten oder Merkmale beziehen, die mit jedem der oben genannten Aufgabenformate überlagert werden können. Zwei Arten von Merkmalen werden hier für die Leistungsmessungen der nächsten Generation besonders in Betracht gezogen: Adaptivität und Lernressourcen.

Erstens hat die digitale Testdurchführung Computergestützte Adaptive Tests (CAT) ermöglicht. Als eine der am meisten erforschten Innovationen in der Testentwicklung (z. B. Wainer et al., 2000) wählen Entscheidungsregeln oder Algorithmen Testaufgaben aus einem Aufgabenpool für einzelne Schüler/innen aus, und obwohl verschiedene Schüler/innen verschiedene Aufgaben oder größere Module in derselben Prüfung bearbeiten können, werden ihre Ergebnisse auf einer gemeinsamen Skala platziert und bleiben vergleichbar. Im Allgemeinen erhöht die Testadaptivität die Effizienz, Genauigkeit und Fairness bei der Entwicklung, Durchführung und Auswertung von Tests, obwohl verschiedene CAT-Designs unterschiedliche Stärken und Schwächen haben (siehe Kasten 7).

KASTEN 7.

COMPUTERGESTÜTZTE ADAPTIVE TESTS: MÖGLICHKEITEN UND HERAUSFORDERUNGEN

Stärken und Schwächen der verschiedenen CAT-Designs

Es wurden bereits verschiedene CAT-Konzepte erforscht und in groß angelegten Tests eingesetzt. Bei einfacheren adaptiven Designs werden die Testaufgaben in Module mit unterschiedlichem Schwierigkeitsgrad eingeteilt und ein Computeralgorithmus leitet die Schüler/innen je nach Leistung in dieses oder jenes Modul. Die Tests können mehrere Stufen umfassen, und die Stufen umfassen mehrere Module (je nach Modul und Testlänge). Verschiedene Algorithmen können verwendet werden, um Verzweigungsentscheidungen zwischen den Stufen zu treffen. Bei diesen Entwürfen erfolgt die Adaption auf der Stufenebene. Andere Konzepte verwenden eine Adaptivität „on-the-fly", bei der die Adaption auf der Ebene der Aufgaben erfolgt (d. h. jede Aufgabe wird auf Grundlage der Leistung einer Person bei früheren Aufgaben auf sie zugeschnitten). Ein Vorteil des einstufigen oder mehrstufigen adaptiven Testens (Multi-Stage Adaptive Testing, MSAT) gegenüber dem adaptiven Testen auf Aufgabenebene besteht darin, dass die Module größere und komplexere Aufgabenformate enthalten können, die ihre eigene interne naturalistische Logik für die in der Aufgabe enthaltenen Items haben. Umgekehrt sind „On-the-fly"-Ansätze, bei denen die Testformen nicht a priori bei der Testentwicklung, sondern zum Testzeitpunkt vom Computer definiert werden, effizient in der Bereitstellung von Aufgaben mit dem jeweiligen Satz an Vorgaben und können eine genauere Schätzung der pro Testzeiteinheit demonstrierten Fähigkeiten liefern. Die Schwäche, Entscheidungen über die nächste Aufgabe ausschließlich auf die Leistung zu stützen, besteht darin, dass dies zu einer geringeren Konstruktabdeckung führen kann und zu einem willkürlichen (statt kohärenten oder thematischen) Weg des Probanden durch die Inhaltsdomäne. Jüngste Fortschritte in Bezug auf CAT können dazu beitragen, dieses Problem zu lösen, indem hybride Messmodelle integriert werden, obwohl diese Designs weit weniger ausgereift sind als ihre gut erforschten Gegenstücke.

Ein anderes CAT-Design passt die Aufgaben auf der Grundlage früherer Entscheidungen oder Handlungen des/der Testteilnehmenden an – ähnlich wie Videospiele auf Basis der Handlungen und Verhaltensweisen der Spieler/innen. Dieser Ansatz hat den Vorteil, dass er die Kontingenzen in realen Problemlösungsumgebungen besser widerspiegelt, und kann, wenn er so gestaltet ist, dass der/die Testteilnehmende eine gewisse Wahlmöglichkeit oder Kontrolle hat, das Engagement erhöhen. Das Ermöglichen einer Adaptivität, die vollständig auf der Wahl des/der Testteilnehmenden basiert, kann jedoch zu konstruktirrelevanter Varianz führen, wenn die Wahl nicht ausdrücklich Teil des Rahmenkonzepts ist. Selbst in Fällen, in denen die Wahlmöglichkeit explizit bewertet wird, können ähnliche Probleme auftreten wie bei spontan adaptiven Modellen ohne ausreichende Beschränkungsmechanismen. Intern adaptive Aufgaben erfordern ebenfalls komplexe Algorithmen. Techniken zur schnellen und effizienten Entwicklung solcher Designs sind noch nicht entwickelt worden, was die Entwicklung und Erprobung dieser Art von Adaptivität kostspielig macht und die Auswertung zum Zweck einer standardisierten Leistungsmessung erschwert. Innovative Leistungsmessungen könnten diese Art von komplexerer, mehrstufiger Adaptivität allerdings integrieren, wenn sie einige der technischen Lösungen übernehmen, die in Videospielen bereits eingesetzt werden, um das Engagement der Spieler/innen aufrechtzuerhalten. Ein Beispiel hierfür wäre der Wechsel zwischen Phasen des Lernens und der *Mastery*.

Quelle: Sabatini et al. (2023), Kapitel 7 in *Innovating Assessments*.

Zweitens erleichtern die digitalen Technologien den Einbezug von Lernressourcen in Leistungsmessungen. Wenn der Schwerpunkt nur darauf liegt, zu messen, wie gut Schüler/innen etwas zu einem bestimmten Zeitpunkt wissen oder können, dann besteht keine Notwendigkeit, Lernressourcen in die Aufgabe einzubeziehen. Innovative Leistungsmessungen könnten jedoch Aussagen darüber treffen, wie Schüler/innen mit authentischen Problemsituationen umgehen, wie sie ihre Problemlösungsstrategien anpassen, wenn sich ihr Verständnis eines Problems verbessert, und wie sie dies tun, indem sie eine Vielzahl von Ressourcen nutzen. In Kasten 8 werden drei Arten von Möglichkeiten beschrieben, die die Integration von Lernressourcen in technologiegestützte Leistungsmessungen ermöglicht.

KASTEN 8.

INNOVATIVE AUFGABENDESIGNS MIT LERNRESSOURCEN

Drei Arten von Hilfestellungen bei technologiegestützten Tests

Lernressourcen bieten vielfältige Möglichkeiten, um zielgerichtetes Verhalten zu ermöglichen. Roll und Barhak-Rabinowitz gruppieren solche Möglichkeiten in drei Kategorien: Experimentieren, explizites Feedback und Informationssuche.

Das Experimentieren ermöglicht es den Lernenden, ihre Ideen zu hinterfragen und darzustellen und sie in einer Weise auszuführen, die Reaktionen aus der Umgebung hervorruft. In Programmierumgebungen können die Lernenden beispielsweise programmieren, kompilieren, ausführen und die Ergebnisse beobachten (umgekehrt werden Programmieraufgaben, bei denen die Lernenden den Code eingeben, ihn aber nicht ausführen können, nach dieser Definition nicht als Lernressourcen betrachtet). Ein weiteres Beispiel sind interaktive wissenschaftliche Simulationen, bei denen die Lernenden Elemente manipulieren und das Ergebnis ihres Experiments beobachten können (z. B. Wieman, Adams und Perkins, 2008). Der Hauptnutzen von Experimentierressourcen ergibt sich aus ihren Reaktionen auf die Handlungen der Lernenden, die oft als situatives Feedback bezeichnet werden (Nathan, 1998; Roll et al., 2014). So passt beispielsweise eine interaktive Simulation für Elektrizität die angezeigte Lichtintensität auf der Grundlage der von den Lernenden eingestellten Spannung an (de Jong et al., 2018; Roll et al., 2018). Situatives Feedback ist implizit und entsteht in der Aufgabensituation selbst, in Übereinstimmung mit der internen Logik der Aufgabe. Das heißt, die Lernenden werden nicht von einem externen, allwissenden Modell beurteilt oder benotet. Stattdessen erhalten sie die Gelegenheit, die relevanten Informationen aus der Reaktion der Umgebung herauszufinden, zu beobachten und zu interpretieren (Nathan, 1998). Die Beobachtung, wie die Lernenden auf situatives Feedback reagieren, kann zur Auswertung ihres Überwachungsverhaltens und der entsprechenden Anpassungen ihrer kognitiven Strategien genutzt werden.

Explizites Feedback bietet den Lernenden eine Evaluierung ihrer Handlungen. Dies kann eine Reihe von Eingaben umfassen, von der Kennzeichnung von Fehlern bis hin zu Erklärungen über die Art des Fehlers oder Vorschlägen für die zukünftige Arbeit (Deeva et al., 2021). Das Feedback kann nach Bedarf (z. B. über eine Schaltfläche „Test") oder automatisch (z. B. nach einer bestimmten Anzahl von Fehlversuchen) ausgelöst werden. Im Gegensatz zu situativem Feedback, das in das Narrativ der Aufgabe integriert ist, ist explizites Feedback extern. Es setzt einen „allwissenden" Agenten oder eine Umgebung voraus, die die Eingaben des Lernenden mit dem gewünschten Ergebnis vergleichen kann.

Die Verwendung von explizitem On-Demand-Feedback bietet eine direkte Messung der metakognitiven Strategien der Lernenden, wie z. B. die Überwachung oder welche Teilziele sie verfolgen (Winstone et al., 2016). Wie bei situativem Feedback zeigen Lernende, die ihre kognitiven Strategien anpassen, indem sie effektiv auf explizites Feedback reagieren, einen produktiven Einsatz von metakognitiven Strategien (z. B. Kinnebrew, Segedy und Biswas, 2017). Hilfestellungen bei der Informationssuche unterstützen die Lernenden, indem sie zusätzliche Informationen über die zu lösende Aufgabe bereitstellen. Zu den Informationsquellen gehören Hinweise (z. B. Aleven et al., 2016), Lehrvideos (z. B. Seo et al., 2021), Arbeitsbeispiele (Ganaiem und Roll, 2022; Glogger-Frey et al., 2015), durchsuchbare Datenbanken usw. Informationsquellen können fix sein (wie in den meisten Tutorials) oder adaptiv (wie in Hinweisen zum spezifischen Problemschritt; VanLehn et al., 2007). Bei der Nutzung von Informationsquellen treffen die Lernenden Entscheidungen darüber, wann sie diese nutzen (z. B. wann sie nach Hinweisen fragen), wie sie sie nutzen (z. B. Navigation durch Videos) und wie sie die Informationen auf die jeweilige Aufgabe anwenden. Effektive und strategische Lernende suchen nach Informationen zum richtigen Zeitpunkt, um ihre eigenen Wissenslücken zu schließen (Seo et al., 2021; Wood, 2001). Daher können die Interaktionen mit Informationsressourcen aussagekräftige Einblicke in die Prozesse des Suchens von Hilfe sowie der Überwachung der Lernenden liefern (Roll et al., 2014).

Für jeden der oben genannten Punkte ist anzumerken, dass das Ermöglichen von Wahlmöglichkeiten im Zusammenhang mit der Bereitstellung von Hilfestellungen ein zusätzliches Konstrukt in der Leistungsmessung darstellt. Die Wahlmöglichkeit muss sich daher ausdrücklich in der Definition des Bereichs widerspiegeln und in die Rückschlüsse auf die Leistung der Schüler/innen einbezogen werden.

Quelle: Roll und Barhak-Rabinowitz (2023), Kapitel 9 in *Innovating Assessments*.

Entscheidungen über die genaue Art und Weise der Unterstützung der Schüler/innen sollten sich an den Zielen der Leistungsmessung orientieren, wie sie im Rahmenkonzept (Kognitionskomponente) festgelegt sind. Wenn beispielsweise die Verwendung von Feedback als konstruktrelevant angesehen wird, sollten intelligente, in die Aufgaben eingebettete Feedback-Mechanismen für die Schüler/innen immer hilfreich sein. Mit anderen Worten: Wenn alle Schüler/innen dieselbe Rückmeldung erhalten, diese aber für einige von ihnen nicht hilfreich ist, sind diese möglicherweise nicht in der Lage, die Zielfähigkeit zu zeigen. Ähnlich verhält es sich, wenn die Wahl des/der Testteilnehmenden konstruktrelevant ist: Vielleicht ist dann ein On-Demand-Mechanismus angemessen. Allerdings kann das Ermöglichen von Wahlmöglichkeiten auch Gelegenheiten ausschließen, solche Verhaltensweisen zu beobachten, so dass es wünschenswert sein kann, auch einige handlungs- oder ereignisgesteuerte Feedbackmechanismen einzubauen.

Eine zentrale Herausforderung bei der Einführung von Lernressourcen in Leistungsmessungen besteht darin, zu entscheiden, welche Auswertungsmodelle angewendet werden sollen. Diese Unterstützungssysteme können den Wissensstand des/der Testteilnehmenden im Verlauf des Tests verändern und so die Leistung der Schüler/innen bei zukünftigen Aufgaben beeinflussen. Noch zu leisten wäre die umfangreichen Forschungsarbeiten, die zu Feedback, Scaffolding und Ressourcen als

Lernmitteln durchgeführt wurden, um Arbeiten im Bereich der psychometrischen Modellierung beim Design von Leistungsmessungen zu ergänzen. Wenn die Testteilnehmenden beispielsweise mehr als eine Chance haben zu antworten (z. B. nachdem sie ein Feedback erhalten haben), könnten die Auswertungsmodelle die richtige Antwort beim ersten Mal höher gewichten als die nachfolgenden Versuche. Alternativ kann es sein, dass das Erreichen einer richtigen Antwort, auch mit Unterstützung, die volle Anerkennung rechtfertigt. Eine enge Zusammenarbeit mit einem Psychometrieteam während des Entwicklungsprozesses des Tests ist entscheidend für das Verständnis der Arten von Inferenzen, die gezogen werden können, und für die Frage, wann solche Merkmale in das statistische Modell integriert werden sollen.

NEUE QUELLEN FÜR EVIDENZ: PRODUKT- UND PROZESSDATEN

Computergestützte Tests erweitern das Spektrum möglicher Evidenzquellen in Leistungsmessungen. Die Palette potenzieller Belege geht weit über die traditionellen Multiple-Choice-Antworten oder konstruierten (schriftlichen) Antworten hinaus, die bei traditionellen Messdesigns, insbesondere bei groß angelegten Tests, vorherrschend waren. Eine wichtige konzeptionelle Unterscheidung in diesem Sinne ist jene zwischen Antwortprodukten und Antwortprozessen und den verschiedenen Arten von Belegen, die diese unterschiedlichen Datenquellen erzeugen (siehe Tabelle 1).

TABELLE 1.
EVIDENZQUELLEN

Produktdaten und Prozessdaten

PRODUKTDATEN	PROZESSDATEN
Verschiedene ausgewählte Antworten (z. B. Multiple Choice, Richtig/Falsch, Drag-and-drop, Hotspot, etc.)	Zeitangaben (z. B. Zeit bei der Aufgabe, Zeit bis zur ersten Handlung, inaktive Zeit)
Schriftliche Antwort	Zwischenzustände der Lösung (d. h. die Zustände vor Einreichen der endgültigen Lösung)
Mündliche Antwort	Aktionsprotokolle (z. B. Nutzung von Hilfestellungen, Tastatureingaben, Mausklicks, Ereignisse)
Performanz (z. B. Erreichen eines Levels in einem Spiel, Simulationszustand, Artefakt)	Physiologische Messungen (z. B. Eye-Tracking-Daten)

Quelle: Sabatini et al. (2023), Kapitel 7 in *Innovating Assessments*.

Antwortprodukte beziehen sich auf die endgültigen Antworten der Schüler/innen auf eine Testaufgabe oder ein bestimmtes Item; Antwortproduktdaten beziehen sich daher in der Regel auf Daten, die aus ausgewählten Antworten (z. B. auf ein Multiple-Choice-Item), kurzen oder längeren schriftlichen Antworten oder dem Endprodukt in einer simulierten oder realen Demonstration einer Leistung resultieren. Antwortprozesse beziehen sich hingegen auf die Denkprozesse,

Strategien und Herangehensweisen der Testteilnehmenden beim Lesen, Interpretieren und Formulieren von Lösungen zu Testaufgaben (Ercikan und Pellegrino, 2017). Antwortprozesse gehen über den kognitiven Bereich hinaus und umfassen Emotionen, Motivationen und Verhaltensweisen (Hubley und Zumbo, 2017). Daten, die potenzielle Belege für diese Prozesse erfassen, können daher als (Antwort-)Prozessdaten verstanden werden, zu denen typischerweise Daten gehören, die Handlungen oder Handlungssequenzen darstellen, Eye-Tracking-Daten und Zeitangaben sowie Daten, die über das spezifische Antwortformat hinausgehen, wie z. B. aufgabenbegleitende Chats und Dialoge mit virtuellen Agenten oder anderen Menschen.

Die einfachste Form von Produktdaten wird durch ausgewählte Antwortformate wie Multiple-Choice-Aufgaben oder Richtig/Falsch-Aufgaben erzeugt, bei denen den Schüler/innen vordefinierte Antworten vorgegeben werden. Diese Antwortformate sind einfacher und kostengünstiger auszuwerten als andere Formate, aber die Testteilnehmenden können die richtige Antwort möglicherweise erraten, und ganz allgemein können diese Formate keinen direkten Beleg der produktiven Kompetenzen liefern.

Andere Formen von Produktdaten (konstruierte Antworten) können diesen Nachweis erbringen, wie z. B. schriftliche Antworten (von einzelnen kurzen Sätzen bis hin zu längeren Aufsätzen), mündliche Antworten oder durch das Anfertigen eines Artefakts oder einer Darstellung (z. B. die Beschäftigung mit einem realistischen Gebäudeentwurf in einer Architekturprüfung oder die Durchführung einer Operation in einem medizinischen Simulator). Dadurch, dass die Schüler/innen eine produktive Aktivität durchführen müssen, sind konstruierte Antworten weniger anfällig dafür, ungerechtfertigterweise Schüler/innen für Rateverhalten zu belohnen, und sind besser geeignet, um Belege für erfolgreiches Lernen und Problemlösen zu erbringen. Sie verlangen von den Testteilnehmenden jedoch auch einen größeren Aufwand und die von ihnen erzeugten Daten können komplexer sein, wenn es darum geht, sie auf zuverlässige und vergleichbare Weise auszuwerten. Zum Beispiel können typische Auswertungsmodelle die Form von Bewertungsrastern oder Erwartungshorizonten annehmen, was jedoch das Design authentischer Aufgaben einschränken kann, indem sie die Art von Antworten verlangen, für die geschulte Bewerter zuverlässige Qualitätsurteile abgeben können.

Fortschritte in der Technologie und Datenanalyse (z. B. linguistische Datenverarbeitung, Spracherkennungssoftware) sind im Begriff, einige dieser Hindernisse zu beseitigen. So können zum Beispiel syntaktische Analysewerkzeuge verwendet werden, um die Struktur der Schülerantworten zu bewerten, und Algorithmen für maschinelles Lernen können trainiert werden, um semantische Ähnlichkeiten zwischen den Schülerantworten und den Lösungen zu erkennen (siehe Hu, Shubeck und Sabatini, 2023; Kapitel 10 in *Innovating Assessments*).

Die Entstehung von Antwortprozessdaten

Neben den Antwortprodukten ist ein herausragender Durchbruch bei technologiegestützten Tests die Eigenschaft, Evidenz aus Antwortprozessen zu generieren. Die Interaktionen der Schüler/innen mit digitalen Testumgebungen können protokolliert werden, um Daten darüber zu erhalten, wie sie bestimmte Prozesse durchführen. Dies kann entscheidend sein, um zu verstehen, was die Schüler/innen beim Lösen einer Aufgabe tun und warum sie es tun. Antwortprozessdaten bieten

die Möglichkeit, diese Handlungen sichtbar zu machen, einschließlich der Frage, wo und wie die Schüler/innen ihre Zeit verbringen und welche Entscheidungen sie in interaktiven und immersiven Umgebungen treffen, was Rückschlüsse auf das Denken der Schüler/innen erleichtert (Ercikan und Pellegrino, 2017).

Prozessdaten können außerordentlich vielfältig sein (z. B. Online-Verhalten, Gestik und Mimik, verbale Interaktion, Augenbewegungen), und jede dieser Datenquellen kann zum Verständnis eines Aspekts der Art und Weise beitragen, wie Testteilnehmende mit Testaufgaben umgehen. In diesem Sinne können Prozessdaten einen Leistungsnachweis darstellen, wenn geeignete Interpretationsmethoden angewandt werden, um gültige Rückschlüsse zu ziehen. Sie können aber auch ein äußerst wertvolles Instrument bei der Validierung von Leistungsmessungen darstellen, indem sie dazu beitragen, zu verstehen, wie verschiedene Schüler/innen mit einer bestimmten Testumgebung umgehen (siehe Ercikan, Guo und Por, 2023; Kapitel 12 in *Innovating Assessments*).

ERNEUERUNG DER *INTERPRETATIONS*-KOMPONENTE: BEOBACHTUNGEN RICHTIG VERSTEHEN

Die vorangegangenen Abschnitte betonten den zunehmenden Konsens darüber, dass sich Leistungsmessung auf das Wesentliche konzentrieren muss, dass zur Messung dieser komplexeren Kompetenzen den Schüler/innen offene und interaktive Testaufgaben in authentischen Kontexten gestellt werden müssen und dass technologiegestützte Leistungsmessungen das Spektrum der belastbaren Belege für das Treffen von Messaussagen erweitern können, einschließlich Datenquellen, die Aufschluss darüber geben können, wie Schüler/innen denken, handeln und lernen, wenn geeignete Interpretationswerkzeuge zur Verfügung stehen beziehungsweise belastbare Argumentationen vorliegen. Hier liegt das dritte Argument für eine innovative Leistungsmessung: Während die Definition von Messkonstrukten komplexer Kompetenzen und die Erfassung neuer Formen von Evidenz mit der Unterstützung von Fachleuten und digitaler Technologie relativ „einfach" ist, ist es weitaus komplizierter, vertretbare Interpretationen der Bedeutung dieser Evidenz vorzunehmen.

Die Herausforderung ergibt sich aus der Tatsache, dass die Interpretationskomponente des Assessment Triangle eigentlich aus zwei Aspekten besteht: der Erhebung von Belegen und der Anhäufung dieser Belege, um eine Aussage über die Kenntnisse, Fähigkeiten oder Einstellungen der Schüler zu treffen. Beides muss argumentierbar sein, einschließlich des Nachweises der Richtigkeit und Präzision der beteiligten Messgrößen und des Ausschlusses alternativer Hypothesen, sowie der Überprüfung, ob die Leistungsmessung fair und angemessen für Teilpopulationen ist. Vor der Dokumentation sollten sowohl die Erhebung als auch die Zusammenstellung der Belege transparent, begründet und gerechtfertigt sein.

EIN PRINZIPIENORIENTIERTER ANSATZ, UM KOMPLEXE DATEN ZU DEUTEN: EVIDENZREGELN UND STATISTIK BEI LEISTUNGSMESSUNGEN

Wie bereits in Abbildung 3 zusammengefasst, sind zwei Komponenten für die Erstellung von Begründungen oder vertretbaren Interpretationen bei groß angelegten Leistungsmessungen erforderlich: Evidenzregeln und Statistik. Diese legen fest, wie den beobachtbaren Variablen Werte zuzuordnen sind und wie die Daten zu Indikatoren oder Skalen zusammengefasst werden sollen.

Evidenzregeln

Evidenzregeln ordnen den Handlungen und Verhaltensweisen der Schüler/innen eine Punktzahl zu. Die Formulierung solcher Regeln ist bei traditionellen und nicht-interaktiven Prüfungen recht einfach, insbesondere wenn Multiple-Choice-Aufgaben verwendet werden: Wenn ein/e Schüler/in eine richtige Antwort auswählt, erhält er/sie Punkte. Bei komplexeren Performanzaufgaben müssen die Merkmale von Arbeitsprodukten oder anderen greifbaren Belegen, die von Expert/innen im Fach mit den Kenntnissen, Fähigkeiten und Einstellungen im jeweiligen Bereich in Verbindung gebracht werden, beschrieben werden. Bei simulations- oder spielbasierten Leistungsmessungen beruhen die Evi-

denzregeln häufig auf der Interpretation von Handlungen und Verhaltensweisen, die als Prozessdaten aufgezeichnet werden (siehe das Beispiel in Kasten 9).

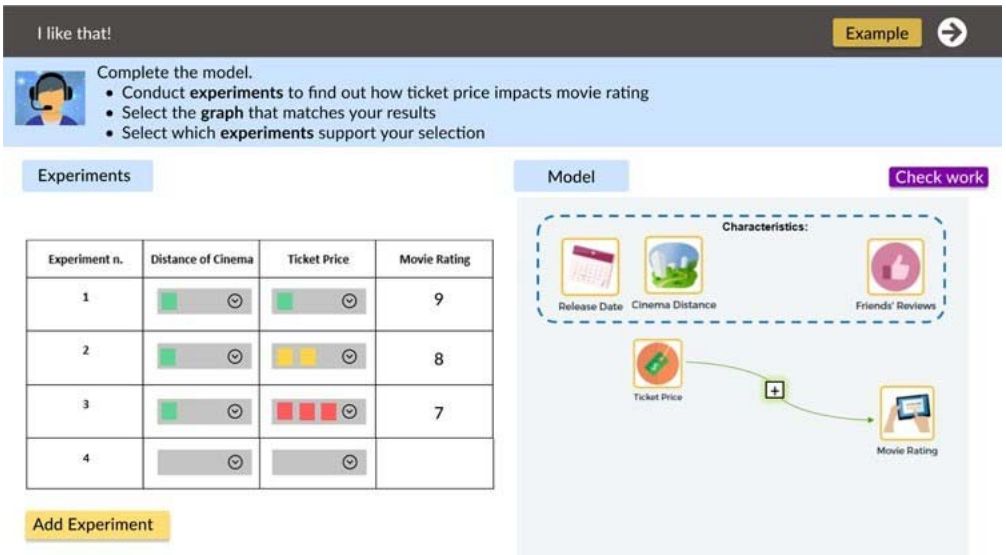
Die Interpretation von Prozessdaten ist jedoch fehleranfällig, da Handlungen in offenen und interaktiven digitalen Umgebungen oft auf unterschiedliche Weise interpretiert werden können. Die Beobachtung, dass ein/e Testteilnehmer/in mit allen Hilfestellungen einer Simulationsumgebung interagiert, könnte beispielsweise als hohes Engagement (d. h. der/die Schüler/in erkundet selbstbewusst Möglichkeiten) oder umgekehrt als hohes Desengagement (d. h. der/die Schüler/in beschäftigt sich nicht sinnvoll mit der Aufgabe) interpretiert werden. Die Definition von Evidenzregeln in diesen Umgebungen erfordert daher: (1) die Rekonstruktion des Raums möglicher Handlungen, die ausgeführt werden können, und deren Einteilung in sinnvolle Gruppen; (2) die Bestimmung des Ausmaßes, in dem Handlungen vom Zustand der Umgebung (und damit von früheren Handlungen) abhängen; und (3) die Verwendung dieser Informationen zur Ermittlung von Sequenzen kontextualisierter Handlungen, die die Beherrschung der angestrebten Kenntnisse, Fähigkeiten und Einstellungen demonstrieren und die in beschreibende Indikatoren oder Punkte umgewandelt werden können.

KASTEN 9.

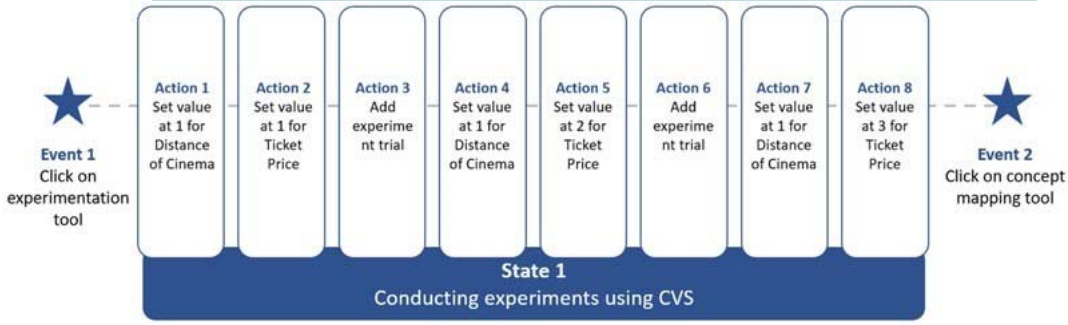
VERWENDUNG VON PROZESSDATEN ALS EVIDENZQUELLEN

Der Fall der „Gefällt-mir"-Einheit in der PISA-Erhebung 2025 „Lernen in der digitalen Welt"

In einer Prototyp-Aufgabe für die PISA-Domäne „Lernen in der digitalen Welt", mit der die Fähigkeit der Schüler/innen, „Experimente durchzuführen und Daten zu analysieren", nachgewiesen werden soll (siehe Abbildung unten), müssen die Schüler/innen ein Tool verwenden, um Experimente durchzuführen, bei denen sie die Variablenkontrollstrategie anwenden, d. h. sie variieren die Werte der unabhängigen Variablen, während alle anderen Variablen konstant bleiben.



Schnittstelle der „Gefällt-mir"-Einheit



Handlungsschritte zur Umsetzung der Variablenkontrollstrategie

Um die Arbeit eines Schülers/einer Schülerin zu bewerten, werden die in den Protokolldaten erfassten Handlungsschritte mit einer Expertenlösung verglichen (Bild oben). Es können Regeln zur teilweisen Punktevergabe entwickelt werden, um Schüler/innen zu erkennen, deren Prozessdaten zeigen, dass sie die Logik der kontrollierten Experimente verstanden haben, aber bei der Ausführung der Strategie einen Fehler gemacht haben (z. B. Prüfung von nur wenigen Werten der unabhängigen Variablen). Wie bei anderen ähnlichen technologiegestützten Aufgaben ist es wichtig, bei der Festlegung der Regeln die Gefahr konstruktirrelevanter Varianz zu berücksichtigen. Ein Beispiel für konstruktirrelevante Varianz in dieser Prototyp-Aufgabe könnte die Unfähigkeit der Schüler sein, die Variablenkontrolle (oder überhaupt ein Experiment) durchzuführen, weil sie nicht in der Lage sind, die Dropdown-Menüs des Experimentiertools zu nutzen.

Quelle: OECD (erscheint in Kürze), PISA-2025-Rahmenkonzept – Lernen in der digitalen Welt (erster Entwurf).

Im Prozess der Definition von Evidenzregeln für komplexe Leistungsmessungen müssen Aufgabendesigns häufig überarbeitet werden, um entweder Möglichkeiten der Hilfestellung hinzuzufügen, um gezielte Handlungen zu erfassen, oder um die Umgebung stärker einzuschränken, um die Bandbreite möglicher Handlungen und Interaktionen zu reduzieren. Mehrfache Iterationsschleifen aus empirischen Analysen und Diskussionen mit Expert/innen im jeweiligen Fachgebiet sind daher für die Identifikation von Evidenz in interaktiven Umgebungen unerlässlich. Oft kombiniert dieser Prozess A-priori-Hypothesen über die Beziehungen zwischen Beobachtungsgrößen und Kenntnissen, Fähigkeiten und Einstellungen mit explorativer Datenanalyse und Daten.

Mislevy et al. (2012) beschreiben dieses Zusammenspiel zwischen Theorie und Entdeckung für eine Testaktivität, die die Konfiguration eines Computernetzwerks beinhaltet. Die Forscher führten eine konfirmatorische Analyse für eine Reihe von Bewertungsregeln durch, die von Expert/innen definiert wurden und die Merkmale der von den Testteilnehmenden eingereichten Arbeitsprodukte berücksichtigten (z. B. ein bestimmter Abschnitt des Netzes wird als „korrekt" betrachtet, wenn Daten von einem Computer zum anderen übertragen werden). Sie ergänzten diese Belege aus Arbeitsprodukten durch die Anwendung von Data-Mining-Methoden auf mit Zeitstempeln versehene Protokolldateien-Einträge. Bei dieser Analyse wurden bestimmte Merkmale wie die

Anzahl der für die Konfiguration des Netzwerks verwendeten Befehle, die Gesamtzeit und die Anzahl der Umschaltvorgänge zwischen Netzwerkgeräten als zusätzliche potenzielle Anhaltspunkte ermittelt, deren Kombination als Maßstab für die Effizienz dienen könnte.

Auswahl eines geeigneten statistischen Modells

Der zweite Aspekt der Interpretationskomponente ist das statistische Modell, das Daten über Aufgaben oder Testsituationen hinweg zusammenfasst, und zwar in Form von aktualisierten Annahmen über Variablen des Schülermodells. Ziel des statistischen Modells ist es, die Beziehung zwischen den beobachteten Variablen (Antworten, Arbeitsprodukte, Handlungssequenzen) und den Kenntnissen, Fähigkeiten und Einstellungen der Schüler in Wahrscheinlichkeiten auszudrücken. Die im Beurteilungsrahmen beschriebenen Modellierungsspezifikationen bieten eine Grundlage für operative Entscheidungen während der Testkonstruktion, z. B. für die Entscheidung, wie viele Aufgaben erforderlich sind, um vertretbare Inferenzen auf der Grundlage der Testergebnisse zu ziehen.

Die einfachsten Messmodelle summieren korrekte Antworten, um Rückschlüsse auf die Kompetenz zu ziehen, während komplexere Messmodelle latente Variablenmodelle wie Item-Response-Modelle (z. B. de Ayala, 2009; Reckase, 2009), diagnostische Klassifikationsmodelle (z. B. Rupp, Templin und Henson, 2010) und Bayes'sche Netze (z. B. Levy und Mislevy, 2004; Conati, 2002) verwenden.

Innovative Leistungsmessungen, die offenes Lernen und Problemlösen simulieren, können wertvolle Erkenntnisse über die Fähigkeiten der Schüler liefern, die jedoch mit bestehenden Messmodellen nur schwer zu erfassen sind. Da die Struktur und die Art der Daten, die in technologieintensiven Aufgaben erhoben werden, zwischen den einzelnen Testteilnehmenden stark variieren können und da die Testaufgaben in offenen und erweiterten Aufgaben voneinander abhängen können, ist es schwierig oder inadäquat, die gleichen psychometrischen Methoden anzuwenden wie für traditionellere Leistungsmessungen (Quellmalz et al., 2012). Diese Erkenntnisse können nur mit Hilfe neuer computergestützter psychometrischer Verfahren voll ausgeschöpft werden. Eine wichtige künftige Herausforderung für innovative Leistungsmessungen besteht darin, das Potenzial computergestützter Methoden für den Umgang mit den reichhaltigeren Daten aus offenen und interaktiven Umgebungen zu verfeinern und nutzbar zu machen und gleichzeitig die Inferenzstärke etablierter psychometrischer Methoden zu bewahren.

EINE GESCHICHTE AUS ZWEI WELTEN: ANSÄTZE DES MASCHINELLEN LERNENS UND EVIDENZBASIERTES DESIGN

Wissenschaftler/innen im Bereich Learning Analytics (LA) und Educational Data Mining (EDM) haben enorme Fortschritte bei der Anwendung von Techniken des maschinellen Lernens (ML) gemacht, um hilfreiche Erkenntnisse aus den Datenströmen zu gewinnen, die in offenen digitalen Lernumgebungen erzeugt werden. Ziel dieser Forschung ist es oft, zu beschreiben, wie Lernende lernen, oder Wege zu finden, Inhalte an einzelne Lernende anzupassen und zu personalisieren. Diese neuen Methoden und die rasanten Fortschritte in der Computertechnologie, die sie unterstützen, bilden die Mittel, um selbst in großem Maßstab Muster im Denken der Lernenden zu erkennen. Nun müssen die Vorteile dieser neuen datengesteuerten Berechnungsalgorithmen genutzt werden, um neue analytische Modelle für Messaussagen zu erstellen,

wobei eine gute Ausrichtung an den grundlegenden Konzepten der Psychometrie beibehalten werden muss.

Das ist nicht so einfach, wie es scheinen mag, denn die beiden Bereiche Psychometrik und Learning Analytics haben ganz unterschiedliche Forschungswege eingeschlagen. Mehr als sechs Jahrzehnte Forschung in Psychometrie und Messmethoden haben gut akzeptierte Verfahren für wichtige Fragen der summativen Leistungsmessung hervorgebracht, darunter Kalibrierung und Schätzung von Gesamtergebnissen, Zuverlässigkeits- und Präzisionsinformationen, Erstellung verschiedener Testversionen, Verlinkung und Äquivalenzierung, adaptive Testverfahren, Überprüfung von Annahmen, Überprüfung der Datenmodellanpassung, differentielle Funktion und Invarianz. Black-Box-Modelle des maschinellen Lernens, wie z. B. Deep Learning mittels künstlicher neuronaler Netze, können sich nicht auf solche Verfahren stützen, weshalb es schwieriger ist, ihnen in Bezug auf Aussagen über die Fähigkeiten von Schüler/innen zu vertrauen – insbesondere dann, wenn es bei diesen Aussagen um viel geht. Sind belastbare Inferenzen aus diesen ML-Modellen möglich, wenn sie nicht in der Lage sind, Gesamtergebnisse zu kalibrieren und zu schätzen, Informationen über Zuverlässigkeit und Genauigkeit zu generieren, Untergruppenanalysen durchzuführen und Verknüpfungen und Gleichsetzungen vorzunehmen?

Auf den ersten Blick scheinen evidenzbasiertes Design von Leistungsmessungen und Data Mining im Bildungsbereich im Widerspruch zueinander zu stehen: Ersteres bezieht sich auf einen prinzipiellen Ansatz für das Design von Aufgabensituationen, die bestimmte Arten von zu bewertenden und zu akkumulierenden Belegen hervorrufen, während sich die zweite Methode auf die Entdeckung aussagekräftiger Muster in den verfügbaren Daten konzentriert. Es ist jedoch möglich, ML-Methoden innerhalb eines prinzipiengeleiteten Verfahrens der Leistungsmessung zu verwenden, bei dem ML-Modelle zusätzliche Informationen über die Testteilnehmenden generieren, die mit Evidenzregeln verknüpft und mit anderen Belegen (z. B. Antworten auf Multiple-Choice-Aufgaben) „aggregiert" werden können, um feinere und belastbarere Aussagen zu treffen.

Die einfache, aber wirkungsvolle Idee hinter diesem Ansatz ist, dass statistische Methoden, die über bewährte Messeigenschaften verfügen, wie z. B. die Item-Response-Theorie (IRT), mit Methoden aus der Lernanalytik erweitert werden können, um den Reichtum der verfügbaren Daten in technologieintensiven Aufgaben voll auszuschöpfen. Die daraus resultierenden aggregierten Belege können mit Hilfe von Standard-Diagnoseverfahren analysiert werden und sind somit „vertrauenswürdiger" für jene, die die Leistungsmessung verwenden. Ein Beispiel für diese Methode unter Verwendung eines mIRT-Bayes-Modells (Scalise, 2017) wird in Kasten 10 vorgestellt. mIRT-Bayes verwendet kleine Bayes'sche Netze, um Scores aus Handlungsmustern zu generieren, und verwendet dann ein multidimensionales IRT-Modell, um Scores zu akkumulieren und Inferenzen zu ziehen.

Diese Möglichkeiten zum Aufbau von Stärken über verschiedene Disziplinen hinweg manifestieren sich im Zusammenhang mit authentischen Technologieaufgaben, wie Simulationen oder Serious Games. Solche Aktivitäten beinhalten viele kleinteilige Erfahrungen, die zu Datenmustern führen, die im Hinblick auf die Anforderungen der Leistungsmessung oft aussagekräftig sind. Ein Avatar, der von einem Schüler gesteuert wird, könnte beispielsweise in einem Raum mit zwei Türen landen; der Schüler muss dann entscheiden, welche Tür geöffnet und was im nächsten Raum getan werden soll. Diese Entscheidungen können

mit einem Modell der Eigenschaften und Fähigkeiten der Schüler/innen verknüpft werden und können so als Evidenz für die Revision bestehender Annahmen über die Verfügbarkeit dieser Eigenschaften und Fähigkeiten dienen. Der Weg in die Zukunft ist die Entwicklung eines Messrahmens, der Perspektiven aus beiden Disziplinen umfasst und der das Design und die Analyse sowohl traditioneller als auch innovativer Leistungsmessungen unterstützt (Mislevy et al., 2012).

KASTEN 10.

ANWENDUNG HYBRIDER MODELLE AUF DIE AUFGABE

„Es gibt einen neuen Frosch in der Stadt"

Das virtuelle Tool zur Leistungsmessung „New Frog" ist eine immersive virtuelle Umgebung mit dem Aussehen und Gefühl eines Videospiels. Alle Teilnehmenden übernehmen die Rolle eines Avatars, der sich in der virtuellen Umgebung bewegen kann. Die Dokumentationsziele dieser Leistungsmessung waren mehrdimensional und beinhalteten wissenschaftliche Exploration und Inquiry (wie in den wissenschaftlichen Standards zu dieser Zeit).

Ein Beispielbild des Tools „New Frog"

In „New Frog" wurden die Schüler/innen aufgefordert, das Problem eines Frosches mit sechs Beinen zu untersuchen. Sie konnten wählen, ob sie verschiedene Frösche zur Exploration des Problems untersuchen wollten, wobei die Wahl an sich weder richtig noch falsch war (es handelte sich also nicht um ein typisches „Item" mit einer vordefinierten Antwort). Muster über die Art und Anzahl der untersuchten Frösche (z. B. solche, die sich in verschiedenen Farmen befanden, zusammen mit Wasserproben aus den Farmen) wurden jedoch als konstruktive Informationen betrachtet, und diese Muster konnten in einem kleinen, aber informativen Bayes-Netz dargestellt werden.

Die Bayes-Netz-Akkumulation fügte dem IRT-Modell beträchtliche Informationen hinzu, erwies sich als den Mustern der naturalistischen Aufgabe angemessen und führte zu einer Verringerung des Standardmessfehlers (Scalise und Clarke-Midura, 2018). Tatsächlich erwiesen sich die von den beiden Bayes-Teilnetzen erzeugten Scores als eines der drei informativsten „Items" in der Aufgabe, was die Passung des Modells in der Studie angeht, obwohl sie aus Daten entwickelt wurden, die ursprünglich verworfen worden waren. Dies ist nicht sonderlich überraschend, wenn man bedenkt, dass es sich bei dem Score um ein Muster salienter Beobachtungen handelte. Allerdings war das andere informativste Item ein wesentlich kostenintensiveres, von Menschen ausgewertetes Constructed-Response-Item. Insgesamt wurde bei der Aufgabe eine feinere Differenzierung der Inferenzen ohne zusätzliche Testzeit oder Auswertungsaufwand ermöglicht, und die Stärken von leistungsschwachen Schüler/innen bei der Durchführung von Forschungsaufgaben traten klarer zu Tage.

Quelle: Scalise, Malcom und Kaylor (2023), Kapitel 8 in *Innovating Assessments.*

DIE GRÜNDE FÜR KOMPLEXERE AUFGABEN UND PRAKTISCHE MÖGLICHKEITEN, SIE IN DER DOKUMENTATION ZU NUTZEN

Die Entwicklung von authentischen Aufgaben, die realen Lern- und Problemlösungsumgebungen nachempfunden sind, wie sie in *Innovating Assessments* beschrieben werden, ist ein zentraler Bestandteil der Validitätsbegründung von Leistungsmessungen der nächsten Generation, welche auf Kompetenzen des 21. Jahrhunderts abzielen, wie beispielsweise kollaboratives Problemlösen, die durch Prozesse definiert sind.

Die Einbeziehung komplexerer und authentischerer Aufgaben in Leistungsmessungen hat auch eine wichtige Signalfunktion. Lehrkräfte, Schüler/innen sowie lokale und nationale politische Entscheidungsträger orientieren sich bei der Zielbestimmung von Unterricht und Kompetenzen an den Aufgabentypen in lokalen, nationalen und internationalen Tests. Daher ist es wichtig, dass die Leistungsmessungen jene Formen von Wissen und Kompetenzen sowie die Arten von Lernerfahrungen widerspiegeln, denen wir in den Klassenzimmern mehr Raum geben wollen. Wenn von den Schüler/innen erwartet wird, dass sie sich die komplexen, mehrdimensionalen Fähigkeiten aneignen, die in der Welt von heute und morgen erforderlich sind, dann sollten sie auch in der Lage sein, diese ihre Fähigkeiten zu zeigen. Die Einbindung von Handlungskompetenz und Relevanz in Leistungsmessungen wird wahrscheinlich auch zu einem höheren Engagement der Schüler/innen beitragen und damit die Wahrscheinlichkeit der Beobachtung dessen, was die Schüler/innen im Rahmen ihrer Möglichkeiten leisten können, erhöhen.

Viele Akteure im Bereich der Leistungsmessung stehen noch immer vor der Frage, was die Entwicklung und Einbeziehung authentischerer Aufgaben für ihre Arbeitspraxis bedeutet. Kosten, Versionierung, Kompatibilität mit Plattformen der Leistungsmessung und andere praktische Erwägungen spielen eine Rolle, insbesondere im Zusammenhang mit groß angelegten Messungen, die wiederholbar und vergleichbar sein sollen. Diese Zwänge halten oft davon ab, komplexe Aufgaben zu einigen wenigen Prototypen zu erstellen. Selbst wenn in das Design komplexer Aufgaben investiert wird, führt die Schwierigkeit bei der Anwendung von Standardmessansätzen mit komplexeren Daten, wie oben beschrieben, oft zu Abkürzungen, die den Wert der Integration authentischer und offener Aufgaben von vornherein stark verringern. Beispielsweise werden manchmal Prozessdaten von der Bereitstellungsplattform erfasst, dann aber nicht im Modell verwendet. Stattdessen wird nur die endgültige Antwort als richtig oder falsch kodiert und bleibt somit die einzige Informationsquelle über die Fähigkeiten der Schüler/innen.

Damit diese komplexere Art von Belegen Einzug in die Praxis halten kann, müssen Leistungsmessungen (zumindest vorerst) eine Mischung aus neueren und älteren Item- und Aufgabentypen enthalten und untersuchen, wie die Belege, die durch verschiedene Arten von Aufgabenformaten und Erfahrungen erzeugt werden, trianguliert werden. Eine solche Triangulation könnte dazu beitragen, ein gemeinsames Verständnis des Wertes innovativer Aufgaben für Inferenzen zu entwickeln, und gleichzeitig diese Inferenzen für die verschiedenen Interessengruppen vertretbarer und „vertrauenswürdiger" machen.

Ein weiterer vielversprechender Weg besteht darin, verschiedene Methoden für verschiedene Arten von Aussagen zu verwenden. So könnten beispielsweise etablierte Messmodelle verwendet werden, um eine Skala zu erstellen, die auf zuverlässige und vergleichbare Weise beschreibt, welche Probleme die Schüler/innen zu lösen in der Lage sind. Learning-Analytics-Methoden könnten dann eingesetzt werden, um eine aussagekräftigere Diagnose der von den Schüler/innen verwendeten Strategien und Prozesse zu erstellen. Dies könnte beispielsweise durch eine Clusteranalyse geschehen, die verschiedene „Typen" von Problemlösenden beschreibt. Beschreibungen der Schüler/innen-Arbeit in den einzelnen Clustern können für Lehrkräfte und Schüler/innen sehr hilfreich sein und veranschaulichen, wie die Kompetenzen des 21. Jahrhunderts in unterrichtsrelevanten Kontexten eingesetzt werden.

INNOVATIVE LEISTUNGS-MESSUNG: AUSBLICK

Innovating Assessments zeigt die Fortschritte auf, die bei der Konzeptualisierung und Operationalisierung entscheidender Aspekte von Leistungsmessungen der nächsten Generation bereits gemacht wurden. Es bietet einen Ausblick darauf, worauf deren Fokus liegen sollte, wie sie aussehen und funktionieren sollten. Damit skizziert es die Karte des zum Erreichen des Ziels zu durchquerenden Geländes sowie einige Zwischenziele auf dem Weg dorthin. Die Karte enthält die wichtigen Konstrukte, Innovationen und Praktiken, die erforderlich sind, um Fortschritte zu erzielen, sowie viele der konzeptionellen und technischen Hindernisse, die zu überwinden sind, um die Vision einer innovativen Leistungsmessung der nächsten Generation zu verwirklichen.

INVESTITIONEN IN LEISTUNGSMESSUNGEN DER NÄCHSTEN GENERATION

Eine Reise, wie sie *Innovating Assessments* anstrebt, kann nicht unternommen werden und wird auch nicht erfolgreich sein, wenn nicht mehrere Formen von Kapital investiert werden. Im Folgenden werden drei besondere Kapitalformen zusammen mit einer Erklärung ihrer Bedeutung betrachtet. Dabei handelt es sich um intellektuelles Kapital, finanzielles Kapital und politisches Kapital. Jede dieser Formen ist notwendig, aber für sich allein nicht ausreichend. Gemeinsam bilden sie das Kapital, das erforderlich ist, um die Theorie und Praxis der Bildungsbewertung voranzubringen und ihren größtmöglichen gesellschaftlichen Nutzen im 21. Jahrhundert hervorzubringen.

INTELLEKTUELLES KAPITAL

Wenn es um Innovationen der Leistungsmessung geht, reichen einzelne Disziplinen oder Fachgebiete nicht aus, um zu erreichen, was zu tun ist. Die bisherigen Fortschritte zeigen, dass die Entwicklung von Leistungsmessungen der nächsten Generation von Natur aus ein multidisziplinäres Unterfangen ist. Experten aus verschiedenen Disziplinen müssen zusammenarbeiten, um Lösungen für die vielen konzeptionellen und fachlichen Herausforderungen zu finden, die bereits bekannt sind oder noch entdeckt werden müssen. Es ist von entscheidender Bedeutung, kreative Menschen mit unterschiedlichem Hintergrund und aus verschiedenen Blickwinkeln für die Entwicklung und Durchführung von Leistungsmessungen zu gewinnen und die Zusammenarbeit zwischen ihnen zu fördern. Synergien zwischen Bewertungsentwicklern, Technologieentwicklern, Lernwissenschaftlern, Domänenexperten, Messexperten, Datenwissenschaftlern, Bildungspraktikern und politischen Entscheidungsträgern müssen gefördert werden.

Da Lernen in soziale Kontexte eingebettet ist und von kulturellen Normen und Erwartungen geprägt ist, kann davon ausgegangen werden, dass die Leistungen in verschiedenen Kulturen unterschiedlich sind.

Die Entwicklung valider Leistungsmessungen, insbesondere für komplexe Fähigkeiten, für die keine etablierten Lernprogressionen verfügbar sind, erfordert multidisziplinäre Teams und Fachwissen. Daher muss bei der Entscheidung, was gemessen, wie es gemessen werden soll und wie die Ergebnisse zu interpretieren und einzusetzen sind, der komplexe soziokulturelle Kontext berücksichtigt werden. Die PISA-Studie 2022 zeigt mit der innovativen Domäne des kreativen Denkens (OECD, 20) beispielhaft, wie wichtig es ist, beim Gestalten der Messung eines komplexen Konstrukts die Schwierigkeiten einer sprach- und kulturübergreifenden Vergleichbarkeit zu berücksichtigen.

Zusätzlich zu den kontext- und kulturbedingten Problemen bei Design und Validierung muss sich die Gemeinschaft jener, die die Tests entwickeln, mit komplexen Fragen auseinandersetzen, z. B. mit dem Design von Aufgaben, die authentische Kontexte simulieren und relevante Verhaltensweisen und Belege elizitieren können, mit der Frage, wie die zahlreichen Datenquellen, die technologiegestützte Tests generieren können, zu interpretieren und zu akkumulieren sind, und mit der Frage, wie Schüler/innen in zunehmend dynamischen und offenen Testumgebungen sinnvoll verglichen werden können. Um diese und verwandte Fragestellungen anzugehen, muss sich die Forschung auf die Modellierung und Validierung komplexer technologiegestützter Leistungen konzentrieren, die vielschichtige Datensätze liefern. Dazu gehört auch die Modellierung von Abhängigkeiten und nicht zufällig fehlenden Daten in offenen und erweiterten Prüfungsaufgaben.

Neue Studien haben gezeigt, dass maschinelles Lernen und KI-Technologie Forschenden dabei helfen können, Lernprozesse besser zu verstehen und zu modellieren (Kleinman et al., 2022), und dass sie Inhaltsexpert/innen dabei unterstützen können, den gesamten Problemlöseprozess von Schüler/innen effizient und effektiv zu annotieren (Guo et al., 2022). Arbeiten dieser Art sind notwendig, um die Erkenntnisse aus kleinen kognitiven Laborstudien zu ergänzen und die Lernforschung voranzutreiben.

Auf pragmatischer Ebene, so Schwartz und Arena (2013) müsse das Design von Leistungsmessungen „demokratisiert" werden, so wie auch der Zugang zur Entwicklung von Videospielen durch die rasche Verbreitung von Online-Communities einfacher geworden sei. Crowdsourcing-Plattformen wie das PILA-Projekt der OECD (OECD, 2023) stellen Entwicklern Modellaufgaben zur Verfügung, die sie anpassen können, und betten Datenerfassungsinstrumente ein, die die Arbeit der Forscher bei der Validierung und Messung vereinfachen. Solche Test-umgebungen könnten die oben erwähnte multidisziplinäre Forschungsarbeit wesentlich erleichtern.

Zusammenfassend lässt sich sagen, dass es eine Vielzahl konzeptionell-theoretischer und praktischer Probleme bei der Zusammenführung von Lern-, Daten- und Messwissenschaften gibt, um zu verstehen, wie die aus komplexen Aufgaben gewonnenen Erkenntnisse am besten mit Modellen und Methoden der künstlichen Intelligenz, des maschinellen Lernens, der Statistik und der Psychometrie analysiert und interpretiert werden können. Eine gemeinschaftliche Auseinandersetzung von Lernwissenschaftlern, Datenwissenschaftlern, Messexperten, Bewertungsentwicklern, Technologieexperten und Bildungspraktikern mit diesen Fragestellungen könnten zu einer neuen Disziplin des Learning Assessment Engineering führen.

FINANZIELLES KAPITAL

Leistungsmessungen für die Anwendung und den Einsatz in einem vernünftigen Umfang zu entwickeln ist ein zeitaufwendiges und kostspieliges Unterfangen. Der Großteil der beträchtlichen Mittel, die derzeit auf nationaler und internationaler Ebene für Programme der Leistungsmessung aufgewendet werden, ist für Design und Durchführung groß angelegter Leistungsmessungen bestimmt, die sich auf traditionelle Disziplinen wie Mathematik, Lesen und Schreiben sowie Naturwissenschaften konzentrieren (z. B. das NAEP-Projekt in den USA und die PISA-Studie der OECD). Die meisten dieser Leistungsmessungen fallen unter die konventionellen Vorgaben für die Aufgabenentwicklung, Durchführung, Datenerfassung, Auswertung und Dokumentation. Dies gilt schon seit geraumer Zeit, obwohl die meisten groß angelegten Programme im Bereich Leistungsmessung inzwischen auf technologiegestützte Aufgabenpräsentation, Datenerfassung und Dokumentation umgestellt wurden. Allerdings war die Nutzung vieler der oben beschriebenen technologischen Möglichkeiten kein besonderes Merkmal dieser Programme.

Die Entwicklung und Validierung von technologiebasierten Aufgaben und Umgebungen ist sehr viel kostspieliger als die Aktualisierung aktueller Leistungsmessungen mittels Erstellung traditioneller Aufgaben, bei welchen Standardaufgabendesigns und -spezifikationen zum Einsatz kommen, die mit Hilfe von Technologie anstelle von Papier und Stift präsentiert werden. Solche neuen Instrumente erfordern beträchtliche Forschungs- und Entwicklungsarbeit in Bezug auf Aufgabendesign, Implementierung, Datenanalyse, Bewertung, Dokumentation und Validierung. Wie bereits angesprochen, muss diese Arbeit von interdisziplinären Gruppen durchgeführt werden, die aus Experten im jeweiligen Fachgebiet, Experten für das Entwickeln von Aufgabenstellungen, Psychometrie und UI-Design sowie Programmierern bestehen. Eine nachhaltige Finanzierung für die notwendige Art von Forschung und Entwicklung ist ein Schlüsselelement auf dem Weg zur Leistungsmessung der nächsten Generation.

Ein wesentliches Hindernis auf diesem Weg ist der Mangel an Beispielen für Bewertungsinstrumente komplexer kognitiver Konstrukte, insbesondere an Beispielen, die nach systematischen Designprinzipien wie dem Evidence-Centered Design entwickelt und dann in der Praxis validiert wurden. Jene Fälle, in denen die Arbeit so weit fortgeschritten ist, dass ihre Validität argumentiert werden kann, sowie für die der Nachweis für eine mögliche Skalierbarkeit auf einen großen Maßstab erbracht wurde, haben nur selten die Forschungs- und Entwicklungsumgebungen verlassen, wo sie als Prototypen entwickelt wurden. Dies gilt selbst für Fälle, die in der Fachcommunity der Forschung und Entwicklung von Leistungsmessungen einen hohen Bekanntheitsgrad erreicht haben. Bedauerlicherweise ist es diesen Arbeiten nicht gelungen, die Art und Weise zu ändern, in der Leistungsmessungen in großem Maßstab konzipiert und durchgeführt werden.

Ebenso wichtig sind Investitionen, um bestehende innovative Bewertungsmethoden zur vollen Reife zu bringen, indem sie in größerem Umfang umgesetzt werden, wenn nachgewiesen ist, dass sie die Herausforderung der Messung der wichtigen Konstrukte wirksam angehen können. Gegenwärtige und zukünftige innovative Konzepte zur Leistungsmessung werden wahrscheinlich in den F&E-Labors verbleiben, es sei denn, es werden Mittel zur Verfügung gestellt, um sie aus den

Labors herauszuholen und in den Bereich der groß angelegten Umsetzung zu bringen, wo ihre Wirksamkeit und ihr Nutzen angemessen evaluiert werden können. Erst dann wird es möglich, sie als Ersatz für die derzeitigen Untersuchungsmethoden einzusetzen.

POLITISCHES KAPITAL

Die derzeitige Praxis der Bildungsbewertung ist stark etabliert, insbesondere die Verwendung groß angelegter standardisierter Leistungsmessungen für das Bildungsmonitoring und politische Entscheidungen. Die Standardisierung umfasst, was gemessen wird, wie es gemessen wird, wie die Daten erhoben und dann analysiert werden und wie die Ergebnisse interpretiert und anschließend präsentiert werden. Es handelt sich hierbei nicht um Zufälligkeiten, sondern um das Ergebnis jahrelanger Arbeit innerhalb einer bestimmten Sichtweise in Bezug darauf, was wir über das Wissen, die Fertigkeiten und die Fähigkeiten eines Menschen wissen wollen und müssen, verbunden mit einer hochentwickelten Technologie der Testentwicklung und -verwaltung, die wiederum mit einer Epistemologie der Interpretation der geistigen Welt verbunden ist, welche in einer aus der physischen Welt abgeleiteten Messmetapher wurzelt.

Es ist schwierig, maßgebliche Veränderungen innerhalb bestehender Systeme vorzunehmen, wenn es bereits gut etablierte Strukturen gibt, die in Praxis und Politik fest verankert sind. Die nötigen Veränderungen erfordern einen starken politischen Willen und eine Vision, die die Menschen dazu ermutigt, über das hinauszudenken, was heute oder auch in naher Zukunft möglich ist. Ohne politischen Willen wird es nicht möglich sein, genügend finanzielles Kapital zu generieren, um das intellektuelle Kapital aufzubauen, das für die Entwicklung und Umsetzung von Leistungsmessungen der nächsten Generation und für sinnvolle Veränderungen in der Bildungsbewertung erforderlich ist.

Das benötigte politische Kapital ist nicht auf die politischen Entscheidungsträger auf Landes- und Bundesebene beschränkt. Es umfasst mehrere Teile der wissenschaftlichen Community in der Entwicklung von Bildungsbewertungen, Messung, Psychometrie und Bildungspraxis. Jede dieser Communities hat festgefahrene Annahmen und Praktiken, wenn es um Leistungsmessung geht. Daher muss sich jede Fachgruppe auf eine Vision der Veränderung einlassen, welche durchaus zu Ergebnissen führen kann, die von Aspekten der derzeitigen Standardverfahren abweichen. Wenn zum Beispiel die Kenntnisse und Fähigkeiten eines Schülers nicht länger als diskret und unabhängig angesehen werden, dann kann ihre Messung die Untersuchung des gesamten interaktiven Verhaltens/Prozesses in adaptiven Lernumgebungen erfordern, die reale Szenarien nachahmen. Unabhängig davon, wohin der Prozess führen wird, müssen diese Fachgruppen zusammenarbeiten, um den politischen Willen und das Kapital aufzubringen, die für die Organisation, Unterstützung und Aufrechterhaltung eines solchen Prozesses erforderlich sind.

INTERNATIONALE GROSS ANGELEGTE LEISTUNGSMESSUNGEN: MÖGLICHKEITEN FÜR INNOVATION IN GROSSEM MASSSTAB

Es liegt auf der Hand, dass noch viel zu tun ist, um die Innovationsagenda im Bereich der Leistungsmessung im Sinne der vorstehenden Ausführungen voranzubringen. Eine der größten Herausforderungen bei der Umsetzung des Wandels besteht darin, dass eine Skalierung erforderlich ist, um zu zeigen, was möglich ist. Die Skalierung vielversprechender Ideen ist von entscheidender Bedeutung, um zu testen, wie tragfähig diese Ideen und Bewertungsansätze sein können, sowie herauszufinden, was nötig ist, um sie in großem Maßstab in die Praxis umzusetzen. Glücklicherweise gibt es einige Beispiele für derartige Bemühungen, die sowohl in Bezug auf die bestehenden Möglichkeiten als auch die noch zu bewältigenden Herausforderungen sehr lehrreich sind.

Internationale Leistungsmessungen dienen im Allgemeinen als Instrumente des Leistungsmonitoring in Bezug auf aktuelle fachliche Standards. Als solche geben diese Studien Aufschluss darüber, was weltweit geschätzt wird, und liefern in großem Maßstab Informationen über die Leistungen von Schüler/innen. Sie stellen auch ein praktisches Beispiel für die Bündelung von intellektuellem, finanziellem und politischem Kapital dar, welches erforderlich ist, um eine innovative, groß angelegte Leistungsmessungsagenda voranzutreiben. So hat sich das PISA-Programm der OECD zum Beispiel vorgenommen, zusätzlich zu den laufenden regulären Bewertungsprogrammen in den Bereichen Mathematik, Lesen und Naturwissenschaften in jeden Messzyklus eine „innovative" Bewertung aufzunehmen. Auf diese Weise hat die OECD die wichtigen Formen von Wissen und Fähigkeiten des 21. Jahrhunderts aufgezeigt, die im Rahmen der Kontrolle breiterer Bildungsziele bewertet werden sollten. Es soll nun kurz auf ein aktuelles Beispiel aus diesem Programm eingegangen werden, um zu veranschaulichen, was die Versuche, innovative Ideen bezüglich der Bewertung des Lernens in die Praxis umzusetzen, bisher gezeigt haben.

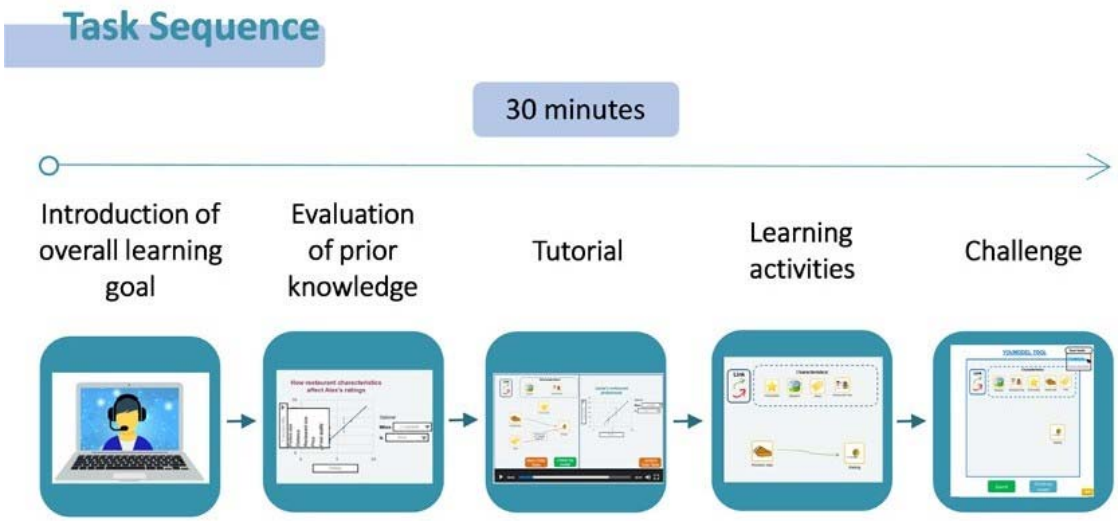
PISA 2025 – LERNEN IN DER DIGITALEN WELT

2025 wird die PISA-Erhebung auch eine Bewertung des Lernens in der digitalen Welt umfassen. Als der PISA-Verwaltungsrat diese neue Entwicklung im Jahr 2020 in Angriff nahm, gab es klare Erwartungen hinsichtlich des Mehrwerts, den sie bringen sollte: Die Länder waren interessiert an vergleichbaren Daten über die Fähigkeit der Schüler/innen in Bezug auf das Lernen und Problemlösen mit digitalen Werkzeugen. Schon vor der weltweiten COVID-19-Pandemie war den Beteiligten klar gewesen, dass digitale Technologien das Bildungswesen erheblich beeinflussen. Dennoch gibt es nicht genügend Informationen darüber, ob die Schüler/innen über die erforderlichen Fähigkeiten verfügen, um mit diesen neuen Werkzeugen zu lernen, und ob die Schulen so ausgestattet sind, dass sie diese neuen Lernmethoden unterstützen können.

Diese politische Forderung bildete den Ausgangspunkt für mehrere Entscheidungen über das Design zukünftiger Leistungsmessungen. Wie bereits erwähnt, stellt eine Bewertung von Lernfähigkeiten andere Anforderungen als eine Bewertung von Wissen. Um effektiv Lernende

von weniger effektiv Lernenden zu unterscheiden, musste der Test den Schüler/innen die Möglichkeit bieten, sich an einer Art von Wissensaufbau zu beteiligen. Mit anderen Worten: Der Test musste als Lernerfahrung strukturiert werden, sodass bewertet werden kann, wie sich das Wissen der Schüler/innen im Laufe des Tests verändert. Folglich hat sich die Struktur der Items vom traditionellen PISA-Format mit einer Reihe von Anregungen und unabhängigen Fragen zu einem neuen Format entwickelt, das als eine Reihe von miteinander verbundenen Lektionen strukturiert ist (Abbildung 6).

Abbildung 6. Aufgabenreihenfolge in der PISA-2025-Domäne „Lernen in der digitalen Welt".



Quelle: OECD (erscheint in Kürze), PISA-2025-Rahmenkonzept Lernen in der digitalen Welt (Entwurf), OECD Publishing, Paris.

Ein virtueller Tutor führt die Schüler/innen durch den Test und erklärt ihnen, wie sie relativ komplexe Probleme mit digitalen Werkzeugen lösen können, die blockbasiertes Programmieren, Simulationen, Datenerfassung und Modellierungsschnittstellen umfassen. Ein interaktives Tutorial mit Videos ist in jede Einheit eingebettet, um den Schüler/innen zu helfen, die Verwendung dieser Werkzeuge zu verstehen und die Unterschiede in der Vertrautheit der Schüler/innen mit bestimmten digitalen Werkzeugen oder Lernumgebungen auszugleichen. Die Schüler/innen lösen dann eine Reihe von Aufgaben, die von Stufe zu Stufe schwieriger werden und sie in die Konzepte und Praktiken einführen, die sie in der Einheit lernen und in der abschließenden und komplexeren „Challenge" einsetzen sollen.

Ein Teil des Beurteilungskonstrukts bezieht sich auf die Fähigkeit der Schüler/innen, selbstgesteuertes Lernen zu betreiben, und erfordert daher die Entwicklung von Maßnahmen wie Monitoring und Anpassung an Feedback sowie die Bewertung von Wissen und Leistung. Um Beobachtungswerte für diese selbstgesteuerten Lernprozesse zu generieren, wurde eine Reihe von Hilfestellungen in die Prüfungsumgebung eingebettet. Im Verlauf des Tests können die Schüler/innen Feedback erhalten, indem sie den Tutor bitten, ihre Arbeit zu überprüfen und zu testen, ob sie die erwarteten Ergebnisse erreicht haben. Sie können sich die Lösungen zu den Übungsaufgaben ansehen, nachdem sie ihre

Antworten eingereicht haben, und für jede Aufgabe können sie Hinweise und Arbeitsbeispiele abrufen, die ihnen bei der Lösung des Problems helfen. Am Ende jeder Aufgabe werden die Schüler/innen aufgefordert, ihre Leistung zu bewerten sowie zu berichten, wie viel sie in die Bearbeitung der Einheit investiert und was sie während der Arbeit empfunden haben. Die Bewertung integriert somit die Annahme, dass komplexe sozio-kognitive Konstrukte besser gemessen werden können, indem den Schülern/innen bei der Bewertung eine Wahl gelassen wird und nicht nur abgefragt wird, wie gut die Schüler/innen Probleme lösen, sondern auch, wie sie dabei vorgehen, um dies zu lernen.

Diese Innovationen sind Antworten auf genau definierte Erfordernisse in Bezug auf Evidenz. Die Leistungsmessung wurde so konzipiert, dass sie Antworten auf drei miteinander verknüpfte Fragen liefert: Welche Arten von Problemen können Schüler/innen im Bereich des computergestützten Entwerfens und Modellierens lösen? Inwieweit sind sie in der Lage, neue Konzepte in diesem Bereich zu erlernen, indem sie Sequenzen von zusammenhängenden, aufeinander aufbauenden Aufgaben lösen? Und inwieweit wird dieses Lernen durch produktive Verhaltensweisen unterstützt, wie z. B. die Entscheidung, bei Bedarf Hilfestellungen zu nutzen oder die Fortschritte bei der Erreichung ihrer Lernziele zu beobachten? Diese Fragen waren bei der Definition des Kognitionsmodells der Bewertung entscheidend sowie beim Design der Aufgaben, welche erforderlich sind, um die notwendigen Beobachtungen machen zu können. Sie leiten auch die Analysepläne, um die Daten auf eine Weise zu interpretieren, die mit den Berichtszwecken der Leistungsmessung übereinstimmt und den komplexen Charakter der Daten berücksichtigt.

Es wird erwartet, dass mehrdimensionale Berichte über die Leistungen der Schüler/innen bei diesem Test erstellt werden, einschließlich Messungen der (1) Gesamtleistung der Schüler/innen bei den Aufgaben (dargestellt in einer Skala, wie bei anderen PISA-Erhebungen); (2) Lernzuwächse, d. h. inwieweit sich das Wissen der Schüler/innen über bestimmte Konzepte und ihre Fähigkeit, bestimmte Tätigkeiten auszuführen, nach dem Training verbessert; und (3) die Fähigkeit, ihr Lernen selbst zu regulieren und ihre Gefühle zu steuern. Diese verschiedenen Messgrößen werden in der Analyse trianguliert, so dass sich beispielsweise ein Teil der Variation im Lernzuwachs aus den Indikatoren für selbstregulierte Lernverhaltensweisen erklärt. Ziel ist es, den politischen Entscheidungsträgern verwertbare Informationen zur Verfügung zu stellen, die sich nicht auf eine Punktzahl und eine Position in einer internationalen Rangliste beschränken, sondern differenziertere Beschreibungen dessen enthalten, was die Schüler/innen leisten können, und aufzeigen, welche Aspekte ihrer Leistung mehr Aufmerksamkeit verdienen.

CODA: ZURÜCK ZU DEN DREI ARTEN VON KAPITAL

Die Entwicklung der 2025-PISA-Domäne „Lernen in der digitalen Welt" wurde erst durch das Zusammenspiel der oben beschriebenen verschiedenen Arten von Kapital ermöglicht. Die politische Unterstützung einer Forschungs- und Entwicklungsagenda durch die PISA-Teilnehmerländer war stark. Die innovative Domäne, die in jeder PISA-Erhebung enthalten ist, wird nun als ein sicherer Raum angesehen, in dem wichtige Innovationen in Bezug auf Aufgabendesign und Analysemodelle erprobt werden können, die dann auf die Hauptdomänen Lesen, Mathematik und Naturwissenschaften übertragen werden können oder die als Anregung für die Entwicklung nationaler Leistungsmessungen dienen können, sobald ihr Wert bewiesen ist.

Der PISA-Verwaltungsrat erkannte die Notwendigkeit mehrfacher Schleifen bei Aufgabendesign und umfassender Validierungsprozesse für die Konzeption und die analytischen Entscheidungen durch kognitive Labors und Pilotstudien an und stellte die finanzielle und politische Unterstützung bereit, die erforderlich war, um mit der Entwicklung des Tests fünf Jahre vor der Hauptdatensammlung zu beginnen. Weitere Ressourcen wurden von Forschungseinrichtungen zur Verfügung gestellt, die den Wert innovativer Bewertungen erkannt haben.

Die Entwicklung der Domäne wurde außerdem von einer Gruppe von Experten mit unterschiedlichem fachlichen Hintergrund gesteuert: Fachexperten arbeiteten Seite an Seite mit Psychometrikern, Lernanalytikern und Experten für UI/UX-Design. Diese gegenseitige Befruchtung war wichtig, um Raum für neue Methoden der Evidenzidentifikation in digitalen Lernumgebungen zu schaffen und gleichzeitig das Hauptziel im Auge zu behalten: nämlich die Erstellung vergleichbarer Metriken, die valide Interpretationen von Leistungsunterschieden zwischen Ländern und Gruppen ermöglichen.

Dieser neue PISA-Test ist nur ein erster Vorstoß in das Gesamtunterfangen der innovativen Bewertungen. Wie in *Innovating Assessments* dargelegt, werden viele neue fachliche und fächerübergreifende Leistungsmessungen benötigt, um eine umfassende Beschreibung der Qualität der Bildungserfahrungen in den verschiedenen Ländern zu erhalten. Darüber hinaus bleiben noch einige Herausforderungen, insbesondere im Bereich der Interpretation des Assessment Triangle. Internationale Foren wie PISA oder die IEA haben die Aufgabe, die politischen Forderungen zu koordinieren und einen Konsens darüber herzustellen, an welchen Schrauben gedreht werden muss und welches die Prioritäten für die nahe Zukunft und darüber hinaus sein sollten. Es ist in hohem Umfang bewiesen, dass eine innovative Messung von pädagogisch und gesellschaftlich bedeutsamen Kompetenzen sowohl wünschenswert als auch möglich ist. Die Belege deuten außerdem darauf hin, dass Kooperation und Zusammenarbeit auf globaler Ebene der beste und einzige Weg sind, um solche Fortschritte zu erzielen.

REFERENZEN

Aleven, V. et al. (2016), „Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems", *International Journal of Artificial Intelligence in Education*, Vol. 26/1, pp. 205-223, https://doi.org/10.1007/s40593-015-0089-1.

Baines, E., P. Blatchford und A. Chowne (2007), „Improving the effectiveness of collaborative group work in primary schools: Effects on science attainment", *British Educational Research Journal*, Vol. 33/5, pp. 663-680, https://doi.org/10.1080/01411920701582231.

Basol, M. et al. (2021), „Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation", *Big Data & Society*, Vol. 8/1, p. 205395172110138, https://doi.org/10.1177/20539517211013868.

Bellanca, J. (2014), *Deeper learning: Beyond 21st century skills*, Solution Tree Press, Bloomington.

Bilal, D. (2000), „Children's use of the Yahooligans! web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks", *Journal of the American Society for Information Science*, Vol. 51/7, pp. 646-665, https://doi.org/10.1002/(SICI)1097-4571(2000)51:73.0.CO;2-A.

Binkley, M. et al. (2011), „Defining Twenty-First Century Skills", in *Assessment and Teaching of 21st Century Skills*, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-007-2324-5_2.

Biswas, G., J. Segedy und K. Bunchongchit (2015), „From design to implementation to practice a learning by teaching system: Betty's Brain", *International Journal of Artificial Intelligence in Education*, Vol. 26/1, pp. 350-364, https://doi.org/10.1007/s40593-015-0057-9.

Brand-Gruwel, S., I. Wopereis und Y. Vermetten (2005), „Information problem solving by experts and novices: Analysis of a complex cognitive skill", *Computers in Human Behavior*, Vol. 21/3, pp. 487-508, https://doi.org/10.1016/j.chb.2004.10.005.

Bransford, J. und B. Stein (1984), The Ideal Problem Solver: *A Guide for Improving Thinking, Learning, and Creativity*, Freeman, New York.

Clark, R. et al. (2008), „Cognitive task analysis", in Spector J. et al. (eds.), *Handbook of Research on Educational Communications and Technology*, Macmillan/Gale, New York, S. 541-551.

Coiro, J. et al. (2019), „Students engaging in multiple-source inquiry tasks: Capturing dimensions of collaborative online inquiry and social deliberation", *Literacy Research: Theory, Method, and Practice*, Vol. 68/1, pp. 271-292, https://doi.org/10.1177/2381336919870285.

Conati, C. (2002), „Probabilistische Bewertung der Emotionen der Benutzer in Lernspielen", *Angewandte Artificial Intelligence*, Vol. 16/7-8, S. 555-575, https://doi.org/10.1080/08839510290030390.

de Ayala, R. (2009), *The Theory and Practice of Item Response Theory*, Guilford Press.

Jong, T. et al. (2018), „Simulations, Games, and Modeling Tools for Learning", in *International Handbook of the Learning Sciences*, Routledge, New York, NY: Routledge, 2018., pp. 256-266, https://doi.org/10.4324/9781315617572-25.

Deeva, G. et al. (2021), „A review of automated feedback systems for learners: Classification

framework, challenges and opportunities", *Computers & Education*, Vol. 162, p. 104094, https://doi.org/10.1016/j.compedu.2020.104094.

Ercikan, K. und M. Oliveri (2016), „In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills", *Applied Measurement in Education*, Vol. 29/4, pp. 310-318, https://doi.org/10.1080/08957347.2016.1209210.

Ercikan, K. und J. Pellegrino (2017), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York, https://doi.org/10.4324/9781315708591.

Foster, N. and M. Piacentini (eds.) (2023), Innovating Assessments to Measure and Support Complex Skills, OECD Publishing, Paris, https://doi.org/10.1787/e5f3e341-en.

Ganaiem, E. und I. Roll (2022), „The effect of different sequences of examples and problems on learning experimental design", Proceedings of the International Conference of the Learning Sciences, Hiroshima, S. 727-732.

Gillies, R. (2016), „Cooperative learning: *Review of research and practice*", *Australian Journal of Teacher Education*, Vol. 41/3, pp. 39-54, https://doi.org/10.14221/ajte.2016v41n3.3.

Gillies, R. und M. Boyle (2010), „Teachers' reflections on cooperative learning: Issues of implementation", *Teaching and Teacher Education*, Vol. 26/4, pp. 933-940, https://doi.org/10.1016/j.tate.2009.10.034.

Glogger-Frey, I. et al. (2015), „Inventing a solution and studying a worked solution prepare differently for learning from direct instruction", *Learning and Instruction*, Vol. 39, pp. 72-87, https://doi.org/10.1016/j.learninstruc.2015.05.001.

Guo, H., Johnson, M., Ercikan, K., Saldivia, L. & Worthington, M. (2022, Juli). Understanding Students' test performance and engagement (Eingeladene Sitzung organisiert/geleitet von K. Ercikan). Internationale Tagung der Psychometrischen Gesellschaft, Bologna, Italien.

Guzdial, M., J. Rick und C. Kehoe (2001), „Beyond adoption to invention: Teacher-created collaborative activities in higher education", *Journal of the Learning Sciences*, Vol. 10/3, pp. 265-279, https://doi.org/10.1207/s15327809jls1003_2.

Hubley, A. und B. Zumbo (2017), „Response Processes in the Context of Validity: Setting the Stage", in *Understanding and Investigating Response Processes in Validation Research, Social Indicators Research Series*, Springer International Publishing, Cham, S. 1-12, https://doi.org/10.1007/978-3-319-56129-5_1.

Irava, V. et al. (2019), „Game-based socio-emotional skills assessment: A comparison across three cultures", *Journal of Educational Technology Systems*, Vol. 48/1, pp. 51-71, https://doi.org/10.1177/0047239519854042.

Jonassen, D. (1992), „What are Cognitive Tools? ", in *Cognitive Tools for Learning*, Springer Ber- lin Heidelberg, Berlin, Heidelberg, S. 1-6, https://doi.org/10.1007/978-3-642-77222-1_1.

Kinnebrew, J., J. Segedy und G. Biswas (2017), "Integrating Model-Driven and Data-Driven Techniques for Analyzing Learning Behaviors in Open-Ended Learning Environments", *IEEE Transactions on Learning Technologies*, Vol. 10/2, pp. 140-153, https://doi.org/10.1109/tlt.2015.2513387.

Kleinman, E. et al. (2022), „Analyzing Students' Problem-Solving Sequences", *Journal of Learning Analytics*, S. 1-23, https://doi.org/10.18608/jla.2022.7465.

Large, A. und J. Beheshti (2000), „The web as a classroom resource: Reactions from the users", *Journal of the American Society for Information Science*, Vol. 51/12, pp. 1069-1080, https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1017>3.0.CO;2-W.

Levy, R. und R. Mislevy (2004), „Specifying and refining a measurement model for a computer-based interactive assessment", *International Journal of Testing*, Vol. 4/4, pp. 333-369, https://doi.org/10.1207/s15327574ijt0404_3.

Lubart, T. (1990), „Creativity and Cross-Cultural Variation", *International Journal of Psychology*, Vol. 25/1, pp. 39-59, https://doi.org/10.1080/00207599008246813.

Messick, S. (1994), „The Interplay of Evidence and Consequences in the Validation of Performance Assessments", *Educational Researcher*, Vol. 23/2, S. 13, https://doi.org/10.2307/1176219.

Mislevy, R. et al. (2012), „Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining", *Journal of Educational Data Mining*, Vol. 4/1, pp. 11-48, https://doi.org/10.5281/zenodo.3554641.

Mislevy, R. und G. Haertel (2007), „Implications of Evidence-Centered Design for Educational Testing", *Educational Measurement: Issues and Practice*, Vol. 25/4, pp. 6-20, https://doi.org/10.1111/j.1745-3992.2006.00075.x.

Mislevy, R. und M. Riconscente (2006), „Evidence-centered assessment design: Layers, concepts, and terminology", in Downing, S. and T. Haladyna (eds.), *Handbook of test development*, Erlbaum, Mahwah, NJ, pp. 61-90.

Nathan, M. (1998), „Knowledge and Situational Feedback in a Learning Environment for Alge- bra Story Problem Solving", *Interactive Learning Environments*, Vol. 5/1, pp. 135-159, https://doi.org/10.1080/1049482980050110.

Niu, W. und R. Sternberg (2001), „Cultural influences on artistic creativity and its evaluation", *International Journal of Psychology*, Vol. 36/4, pp. 225-241, https://doi.org/10.1080/00207590143000036.

OECD (in Vorbereitung), PISA 2025 Learning in the Digital World assessment framework (Entwurf), OECD Publishing, Paris

OECD (2022), *Thinking Outside the Box: The PISA 2022 Creative Thinking Assessment*, https://issuu.com/oecd.publishing/docs/thinking-outside-the-box (Zugriff am 4. März 2023).

OECD (2013), *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, OECD Publishing, Paris, http://www.oecd-ilibrary.org/docserver/download/9113021e.pdf?expires=1511446761&id=id&accname=guest&checksum=18A9C-C493392BE9A918508D9929D29A3.

Pellas, N. et al. (2018), „Augmenting the learning experience in primary and secondary school education: a systematic review of recent trends in augmented reality game-based learning", *Virtual Reality*, Vol. 23/4, pp. 329-346, https://doi.org/10.1007/s10055-018-0347-2.

Pellegrino, J., N. Chudowsky und R. Glaser (2001), *Knowing What Students Know:* The Science and Design of Educational Assessment, National Academy Press.

Pellegrino, J., L. DiBello und S. Goldman (2016), „A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments", *Educational Psychologist*, Vol. 51/1, pp. 59-81, https://doi.org/10.1080/00461520.2016.1145550.

Pellegrino, J. und M. Hilton (2012), *Education for life and work: Entwicklung von übertragbaren Kenntnissen und Fähigkeiten im 21. Jahrhundert*, https://doi.org/10.17226/13398.

Quellmalz, E. et al. (2012), „21st century dynamic assessment", in Mayrath, M. et al. (eds.), *Technology-Based Assessments for 21st Century Skills*, Information Age Publishing, http://www.simscientists.org/downloads/Chapter_2012_Quellmalz.pdf.

Raphael, C. et al. (2009), „Games for civic learning: A conceptual framework and agenda for research and design", *Games and Culture*, Vol. 5/2, pp. 199-235, https://doi. org/10.1177/1555412009354728.

Reckase, M. (2009), *Multidimensional Item Response Theory*, Springer, New York, https://doi.org/10.1007/978-0-387-89976-3.

Roll, I. et al. (2018), „Understanding the impact of guiding inquiry: the relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning", *Instructional Science*, Vol. 46/1, pp. 77-104, https://doi.org/10.1007/s11251-017-9437-x.

Roll, I. et al. (2014), „Tutoring Self- and Co-Regulation with Intelligent Tutoring Systems to Help Students Acquire Better Learning Skills", in Sottilare, R. et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems: Volume 2 Instructional Management*, US Army Research Laboratory, Orlando, S. 169-182.

Roozenbeek, J. und S. van der Linden (2018), „The fake news game: Actively inoculating against the risk of misinformation", *Journal of Risk Research*, Vol. 22/5, pp. 570-580, https://doi.org/10.1080/13669877.2018.1443491.

Rupp, A., J. Templin und R. Henson (2010), *Diagnostic Measurement: Theory, Methods, and Applications*, Guilford Press, New York.

Scalise, K. (2017), „Hybrid Measurement Models for Technology-Enhanced Assessments Through mIRT-bayes", *International Journal of Statistics and Probability*, Vol. 6/3, S. 168, https://doi.org/10.5539/ijsp.v6n3p168.

Scalise, K. und J. Clarke-Midura (2018), „The many faces of scientific inquiry: Effectively measuring what students do and not only what they say", *Journal of Research in Science Teaching*, Vol. 55/10, pp. 1469-1496, https://doi.org/10.1002/tea.21464.

Schwartz, D. und D. Arena (2013), *Measuring what matters most: Choice-based assessments for the digital age*, The MIT Press, Cambridge, Massachusetts.

Seo, K. et al. (2021), „Active learning with online video: The impact of learning context on engagement", *Computers* & Education, Vol. 165, S. 104132, https://doi.org/10.1016/j.compedu.2021.104132.

Sternberg, R. (2013), „Intelligence", in Freedheim, D. and I. Weiner (eds.), *Handbook of Psychology: History of Psychology*, John Wiley & Sons, Hoboken, pp. 155-176.

Toulmin, S. (2003), *The Uses of Argument*, Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511840005.

Urban, A., C. Hewitt und J. Moore (2018), „Fake it to make it, media literacy, and persuasive design: Using the functional triad as a tool for investigating persuasive elements in a fake news simulator", *Proceedings of the Association for Information Science and Technology*, Vol. 55/1, pp. 915-916, https://doi.org/10.1002/pra2.2018.14505501174.

van der Linden, S., J. Roozenbeek und J. Compton (2020), „Inoculating against fake news about COVID-19", *Frontiers in Psychology*, Vol. 11/566790, pp. 1-7, https://doi.org/10.3389/ fpsyg.2020.566790.

VanLehn, K. et al. (2007), *What's in a step? Auf dem Weg zu allgemeinen, abstrakten Repräsentationen von Tutoring-Systemprotokolldaten*, Springer.

Voogt, J. und N. Roblin (2012), „A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies", *Journal of Curriculum Studies*, Vol. 44/3, pp. 299-321, https://doi.org/10.1080/00220272.2012.668938.

Wainer, H. et al. (2000), *Computerized Adaptive Testing*, Routledge, https://doi.org/10.4324/9781410605931.

Wieman, C., W. Adams und K. Perkins (2008), „PhET: Simulationen zur Verbesserung des Lernens", *Science*, Vol. 322/5902, S. 682-683, https://doi.org/10.1126/science.1161948.

Winstone, N. et al. (2016), „Supporting Learners` Agentic Engagement With Feedback: A Sys- tematic Review and a Taxonomy of Recipience Processes", *Educational Psychologist*, Vol. 52/1, pp. 17-37, https://doi.org/10.1080/00461520.2016.1207538.

Wolf, S., T. Brush und J. Saye (2003), „Using an information problem-solving model as a metacognitive scaffold for multimedia-supported information-based problems", *Journal of Research on Technology in Education*, Vol. 35/3, pp. 321-341, https://doi.org/10.1080/1539152 3.2003.10782389.

Wood, D. (2001), „Scaffolding, Contingent Tutoring and Computer-supported Learning", *Inter-Nationale Zeitschrift für Künstliche Intelligenz im Bildungswesen*, Bd. 12, S. 280-293.

Ursprünglich auf Englisch veröffentlicht von

Gefördert von