Marc Rodemer[1] [1] Radboud University Nijmegen, NL
Peter Wulff[2] [2] PH Ludwigsburg, DE
Alexander Krause[3] [3] Leibniz Universität Hannover, DE
Stefan Küchemann[4] [4] LMU München, DE

KI-gestütztes Assessment zwischen Algorithmus und Urteil: Qualitätsdimensionen, Prozessmodell und Perspektiven

Das Symposium *KI Assessment: zwischen Algorithmus und Urteil* bündelt vier Arbeiten zu generativer KI und *Learning Analytics* im naturwissenschaftsdidaktischen Assessment in der Physik und Chemie. Ausgehend von diesen Beiträgen wird ein Ordnungsrahmen entwickelt, der Qualitätsdimensionen, Prozessschritte und die Rolle des menschlichen Urteils integriert. Insgesamt ergänzt KI professionelles Urteil, sie kann es jedoch nicht ersetzen.

Einleitung und Zielsetzung

Generative KI (Zhao et al., 2024) und *Learning Analytics* (Zhang et al., 2023) verändern Assessments in den Naturwissenschaftensdidaktiken (Circi et al., 2023). Lehrpersonen können mit KI schneller Aufgabenentwürfe erzeugen und in großen Kohorten individuelles Feedback organisieren. Gleichzeitig obliegen die zentralen Validierungsarbeiten bei der fachlichen und fachdidaktischen Expertise (Ahmed et al., 2025; Chan et al., 2025; Chatzichristofis, 2025).
Der erste Beitrag (Küchemann et al.) behandelt KI-gestützte Konzeptaufgaben in der Physik und knüpft an etablierte Strukturen an (B1). Der zweite Beitrag (Rodemer et al.) fokussiert KI-erstellte Aufgaben in der Elektrochemie mit Blick auf Kompetenzstufen und sprachliche Zugänglichkeit (B2). Der dritte Beitrag (Wulff et al.) untersucht, wie gezielte Instruktionen an große Sprachmodelle die Qualität von Assessment und Feedback beeinflussen (B3). Der vierte Beitrag (Krause et al.) analysiert die auf einer webbasierten Lernplattform erfassten Logfiles bei Stöchiometrieaufgaben und leitet daraus Perspektiven für adaptive Unterstützung ab (B4).
Im Folgenden sollen anhand der Beiträge drei Ziele verfolgt werden:

1. Beschreibung von Qualitätsdimensionen eines KI-gestützten Assessments.
2. Ableitung eines Prozessmodells, welches die Aufgabenbereiche zwischen KI und Mensch sichtbar macht.
3. Diskussion von Perspektiven für Praxis und Forschung

Qualitätsdimensionen

Die Qualität KI-gestützter Assessments lässt sich entlang weniger, eng verknüpfter Dimensionen prüfen. Reliabilität und Dimensionalität betreffen die Stabilität und inhaltliche Struktur der gemessenen Kompetenz. Es zeigt sich, dass KI anschlussfähige Aufgaben generieren kann, die Stabilität der Messung jedoch nur trägt, wenn Konstruktbindung und Selektion konsequent mitlaufen. Verlässlichkeit entsteht aus der Verbindung fachdidaktischer Modellierung und empirischer Prüfung, nicht aus der Generierung allein (B1, B2).
Validität verlangt, dass Aufgaben das intendierte Konstrukt erfassen und Interpretationen belastbar sind (Downing & Haladyna, 1997). Gute Modellfits reichen ohne inhaltliche Rückbindung nicht aus. Stufenbeschreibungen müssen mit empirischen Verortungen abgeglichen werden, damit Fehlinterpretationen ausbleiben. Sprachliche Entscheidungen als Teil der Validität, verlangen eine eigenständige Zugänglichkeitsprüfungen (B2).

Kognitive Passung entsteht, wenn Aufgaben typische Denkwege und Fehlkonzepte adressieren. Die Qualität von Kompetenzmodellen nimmt durch Evidenzen aus realen Bearbeitungsprozessen zu. Prozessdaten können diese Passung sichtbar machen. So wächst inhaltliche Präzision und es entstehen zugleich Anknüpfungspunkte für spätere Rückmeldungen an Lernende (B1, B4).
Reproduzierbarkeit und Nachvollziehbarkeit betreffen die Stabilität von Modellausgaben und den Umgang mit Prompting. Zielgerichtete Instruktionen und eine dokumentierte Prompt-Governance erhöhen die Kohärenz und machen Ergebnisse replizierbar. Governance ist hier keine Ergänzung, sondern eine Voraussetzung für verlässliche Entwicklung und Pflege von Items sowie Feedbackregeln im laufenden Betrieb (B3).
Fairness verläuft quer zu allen Dimensionen. Sie betrifft sprachliche Hürden, Vorwissen und Strategien. Systematisches Monitoring, transparente Kriterien und definierte Korrekturwege sichern, dass Abweichungen erkannt und behoben werden, bevor sie in die Interpretation von Leistungen eingreifen (B2, B3, B4).
In Summe entsteht Mehrwert durch das Zusammenspiel aus fachlicher Begutachtung, psychometrische Evidenz, prozessdatenbasierte Prüfung und dokumentierte Governance. Wo Diskrepanzen auftreten, greifen definierte Schleifen zur Korrektur und Weiterentwicklung.

Prozessmodell für KI-gestütztes Assessment
Das folgende Modell leitet sich aus den gebündelten Qualitätsdimensionen ab und ordnet die Rollen von KI und Mensch entlang eines Entwicklungs- und Betriebspfads (Abbildung 1).

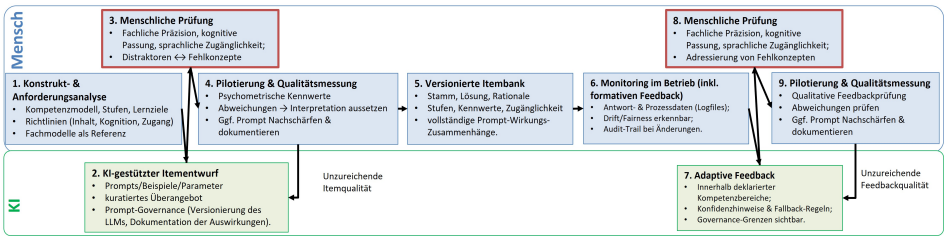

Abbildung 1: Prozessmodell für KI-gestütztes Assessment

Am Anfang steht die Konstrukt- und Anforderungsanalyse. Teams legen das Kompetenzmodell fest, bestimmen Facetten und interpretierbare Stufen und klären den Verwendungszweck (formativ/ summativ). Die Leitplanken aus Inhaltsvalidität, kognitiver Passung und Zugänglichkeit werden hier verankert (B1, B4).
Es folgt der KI-gestützte Itementwurf. Prompts, Beispiele, formale Vorgaben und kontrollierte Parameter sorgen für ein ausreichend breites Spektrum. Ziel ist ein kuratierbares Überangebot, kein sofort perfektes Item. Prompt-Governance mit Versionierung und Dokumentation schafft die Grundlage für Replikation und spätere Revisionen. Gezielte Instruktionen erhöhen die Kohärenz spürbar und senken den nachgelagerten Aufwand, ohne ihn zu ersetzen (B1, B3).
Die menschliche Prüfung bündelt fachliche Präzision, kognitive Passung und sprachliche Zugänglichkeit. In dieser redaktionellen Phase werden Distraktoren an Fehlkonzepte rückgebunden, sprachliche Entscheidungen verantwortet und intendierte Stufen nachvollziehbar dokumentiert. Dieser Schritt übersetzt fachdidaktische Prinzipien in überprüfbare Befunde, die eine tragfähige Interpretationen ermöglichen (B1, B2).
Pilotierungen mit der Zielgruppe klären psychometrische Eigenschaften und die Verortung von Items auf der Kompetenzskala, d.h. ob intendierte und empirische Stufen

zusammenpassen. Stufenbeschreibungen, Distraktoren und Auswahlkriterien werden nachgeschärft, bevor Aufgaben in den Regelbetrieb übergehen (B2).
Tragfähige Items werden in eine versionierte Bank überführt. Neben Stamm und Lösung gehören Rationale, Stufenetiketten, Kennwerte, Hinweise zur Zugänglichkeit sowie der vollständige Prompt in die Metadaten. Die Bank ermöglicht Pflege, Replikation und transparente Rückkopplungen zwischen Entwicklung und Betrieb zugleich (B3).
Im laufenden Einsatz stützt sich das Monitoring auf Antwort- und Prozessdaten. Logfiles zeigen Lösungsstrategien und Lernschwierigkeiten, Drifteffekte werden sichtbar, Fairnessfragen adressierbar. Auffällige Aufgaben werden gesperrt, überarbeitet oder neu kalibriert. Das Erfassen von Prozessdaten bildet zudem die Grundlage für präzises individualisiertes Feedback, welches anhand von Ergebnissen nur bedingt umsetzbar ist (B4).
Adaptive Rückmeldungen entstehen auf dieser Basis als letzter Baustein. Sie bleiben innerhalb deklarierter Kompetenzbereiche und verfügen über Fallback-Regeln, wenn Unsicherheit steigt. Wo Modelle verlässlich sind, übernehmen sie Teile der Formulierung. Wo sie es nicht sind, greifen klare Regeln und vorbereitete Hinweise. Governance sorgt dafür, dass diese Grenze sichtbar bleibt und Entscheidungen wiederholbar sind. So wird die Balance zwischen Algorithmus und Urteil praktisch handhabbar und verantwortlich gestaltet (B3).

Praxis- und Forschungsperspektiven: nächste Schritte

Für die Praxis zählt ein durchgehender Fluss vom Konstrukt zur Rückmeldung. Teams klären zunächst Kompetenzen, erzeugen danach in kontrollierten Prompt-Batches kuratierbare Items und prüfen sie fachlich sowie sprachlich. Pilotierungen sichern Kennwerte und klären per Anchoring die Stufenpassung. Tragfähige Aufgaben wandern versioniert in die Bank und werden über Antwort- und Prozessdaten überwacht. Adaptive Rückmeldungen werden dort eingesetzt, wo stabile Strategieprofile vorliegen und klare Guardrails greifen (B1, B2, B3, B4).
Für die Forschung zeichnen sich mehrere Anschlusslinien ab. Es braucht robuste Verfahren zur Kalibrierung von Kompetenzen und Prozessdaten. Distraktoren-Engineering sollte Fehlvorstellungen nutzen und Plausibilität automatisch prüfen. Metriken für Feedback-Validität und Kohärenz über Prompt- und Modellvarianten sind zu entwickeln, idealerweise mit Konfidenzberichten, die Entscheidungen im Betrieb stützen. Schließlich lohnt es sich, Prozess-zu-Feedback-Pipelines so zu gestalten, dass Logfile-basierte Strategieprofile in gestufte Scaffolds überführt und in Wirksamkeitsstudien überprüft werden. Diese Forschungslinien sind eng mit den im Text angelegten Systemen und Prinzipien verschaltet und lassen sich in laufende Entwicklungsumgebungen integrieren.

Schlussfolgerung

KI kann Assessment in Physik und Chemie skalieren und individualisieren. Die Beiträge zeigen zugleich, dass ohne fachliche Validierung, kognitive Passung und geprüfte Zugänglichkeit kein belastbarer Einsatz zu erwarten ist. Ein Ordnungsrahmen, der Qualitätsdimensionen, Prozessschritte, Governance und Betrieb zusammendenkt, macht die Balance zwischen Algorithmus und Urteil praktisch handhabbar. Die Kombination aus Prompt-Governance, Item- und Promptbanking, evidenzbasierter Psychometrie und fachlich zugeschnittenen Learning-Analytics liefert die Bausteine, um das Potenzial von KI zu heben und Risiken kontrolliert zu halten. KI kann somit als wertvolle Ergänzung zum Assessment verstanden werden, bei der menschliche Expertise weiterhin unverzichtbar bleibt.

Literatur

Ahmed, A., Kerr, E. & O'Malley, A. (2025). Quality assurance and validity of AI-generated single best answer questions. *BMC Medical Education*, *25*(1), 300.

Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K. & Sze, G. K. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, *8*, 100344.

Chatzichristofis, S. A. (2025). Reframing AI in education: A pedagogical opportunity, not a technocratic threat. *Journal of Research in Science Teaching*, online first.

Circi, R., Hicks, J. & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*.

Downing, S. & Haladyna, T. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 1, 61-82.

Zhao, J., Chapman, E. & Sabet, P. G. (2024). Generative AI and educational assessments: A systematic review. *Education Research and Perspectives, 51*, 124-155.

Zhang, K., Yılmaz, R., Ustun, A. B. & Yılmaz, F. G. K. (2023). Learning analytics in formative assessment: A systematic literature review. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(Özel Sayı), 359-381.