MedienPädagogik
Zeitschrift für Theorie und Praxis der Medienbildung
www.medienpaed.com **ISSN 1424-3636**

Themenheft 71: Mediendidaktik 2025 – Tagungsband der 2. Tagung des Arbeitskreises Mediendidaktik der DGfE-Sektion Medienpädagogik
Herausgegeben von Barbara Getto, Tobias M. Schifferle, Josef Buchner, Elke Höfler, Michael Kerres, Franco Rau und Karsten D. Wolf

Personalisierung von Prüfungen durch ChatGPT-4o?

Argumentative Validität von personalisierten Prüfungen im Zeitalter generativer Künstlicher Intelligenz

Lhea Reinhold[1] und Katja Buntins[2]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] Universität Duisburg-Essen

Zusammenfassung
Generative Künstliche Intelligenz (genKI) kann Lern- und Prüfungsprozesse individualisieren – je nach Einsatz kann sie dabei erweiternd oder transformierend wirken (Puentedura 2006). Zu überprüfen gilt, inwiefern solche personalisierten Prüfungen valide sind und in welchem Masse diese das Verhältnis von Standardisierung und Personalisierung in prüfungsbezogenen Prozessen verändern. Vor diesem Hintergrund untersucht die empirische Studie die Validität von zwei personalisierten Prüfungen – eine kontextualisierte, textbasierte Prüfung (erweiternd) und die Nutzung eines Prüfungsbots während der Prüfungssituation (transformierend) – anhand einer argumentativen Evidenzkette (St-Onge et al. 2017; Huggins-Manley et al. 2022), strukturiert durch die prüfungsbezogenen Prozesse scoring, administration, evaluation, decision und impact (Sinharay et al. 2025). In einem Kontrollgruppendesign (n = 18) erhielten Expert:innen u. a. aus dem Bereich Angewandte Künstliche Intelligenz entweder die von einer Lehrkraft entwickelte standardisierte Prüfung oder eine davon abgeleitete, von genKI entwickelte kontextualisierte Prüfung, die beide im Kurs «KI-Management» eines Weiterbildungsinstituts eingesetzt werden könnten. Die Ergebnisse zeigen, dass der erweiternde Personalisierungsgrad die Validität nicht signifikant beeinträchtigt oder aufwertet; ein signifikanter authentizitätsfördernder Effekt zeigt sich bei der kontextualisierten Prüfung. Für den Einsatz eines Prüfungsbots gilt: Validität ist dann gefährdet, wenn die Eigenleistung der lernenden Person nicht mehr ersichtlich ist; vorteilig wirkt der Prüfungsbot beim Geben von Feedback während und nach der Prüfung. Somit könnten Lern- und Prüfungsprozesse weiter ausdifferenziert und die Konzepte Standardisierung und Personalisierung als komplementär verstanden werden (Bates et al. 2019).

Reinhold, Lhea, und Katja Buntins. 2026. «Personalisierung von Prüfungen durch ChatGPT-4o? Argumentative Validität von personalisierten Prüfungen im Zeitalter generativer Künstlicher Intelligenz». *MedienPädagogik* 71 (Mediendidaktik 2025): 233–263. https://doi.org/10.21240/mpaed/71/2026.08.16.X.

Personalization of an Assessment Using ChatGPT-4o? Argumentative Validity of a Personalized Assessment in the Age of Generative Artificial Intelligence

Abstract

Generative Artificial Intelligence (genAI) has the potential to individualize learning and assessment processes – depending on its mode of application, genAI can have either an augmentative or a transformative effect (Puentedura 2006). A central question is to what extent such personalized assessments can be considered valid, and to what degree they alter the relationship between standardization and personalization within assessment-related processes. Against this backdrop, the empirical study investigates the validity of two forms of personalized assessment – one contextualized, text-based assessment (augmentative) and the use of an assessment bot during the assessment situation (transformative) – through an argumentative chain of evidence (St-Onge et al. 2017; Huggins-Manley et al. 2022) structured according to the assessment processes of scoring, administration, evaluation, decision, and impact (Sinharay et al. 2025). Using a control group design (n = 18), experts, including specialists in Applied Artificial Intelligence, received either a teacher-developed standardized assessment or a genAI-developed contextualized assessment derived from it, both of which could be implemented in the course «AI Management» at a continuing education institute. The findings indicate that the augmentative level of personalization neither significantly impairs nor enhances validity; however, a significant effect promoting authenticity was observed for the contextualized assessment. Regarding the use of an assessment bot, validity appears compromised when the learner's individual contribution becomes indistinguishable, yet the bot proves advantageous for providing feedback during and after the assessment. Consequently, learning and assessment processes could be further differentiated, and the concepts of standardization and personalization could be understood to be complementary (Bates et al. 2019).

1. Einleitung

Der gesetzeskonforme Einsatz generativer Künstlicher Intelligenz (genKI) in Prüfungssituationen ist vielfältig: genKI kann bei der Erstellung von Prüfungsaufgaben (Chan et al. 2025), bei der Korrektur von Leistungen (Reinhold und Händel 2025) und beim Geben von Feedback assistieren (Ali et al. 2023). Im besten Fall unterstützen diese Einsatzmöglichkeiten die prüfungsbezogenen Tätigkeiten einer Lehrkraft. Mit dem Ziel, die Vorteile der disruptiven Technologie zu nutzen (Alier et al. 2024), könnten Prüfungen sogar erweitert oder transformiert werden (Puentedura 2006).

Eine solche Möglichkeit stellt die Personalisierung von Prüfungen dar: In einem nachhaltigen Lernprozess ergänzen sich lern- und prüfungsbezogene Prozesse (Boud 2000; Boud und Soler 2016). So müsste auf einen individualisierten Lernprozess ein

individualisierter Prüfungsprozess folgen. An dieser Stelle kann ein Spannungsverhältnis entstehen: Hat eine Prüfung die standardisierte Bewertung der Lernleistung und somit die Vergleichbarkeit von Ergebnissen zum Ziel (Schaper 2021), gilt es, die individualisierte Prüfungssituation im Hinblick auf ihre Validität zu hinterfragen, da Validität als primäres Gütekriterium die Übereinstimmung der Prüfungsaufgabe mit der tatsächlich zu prüfenden Kompetenz sichert (American Psychological Association et al. 1999).

Vor diesem Hintergrund untersucht die explorative Studie, inwiefern eine Prüfung personalisiert *und* valide sein kann. Die Fragestellung wird im Kontext der geförderten Weiterbildung in Deutschland untersucht, indem eine von einer Lehrkraft entwickelte standardisierte, textbasierte Prüfung in zwei Personalisierungsgraden variiert wird. Kriterial ausgewählte Expert:innen bewerten mithilfe von Fragebögen die Validität dieser Personalisierungsgrade, sodass u. a. das Verhältnis von Standardisierung und Personalisierung erschlossen werden kann.

2. Theoretische Konzepte

2.1 Lernen und Prüfen im Zeitalter genKI

Mit dem Zugang zu genKI, z. B. ChatGPT, verändern sich lern- und prüfungsbezogene Prozesse (Nicola-Richmond et al. 2025). Sowohl die Lerninhalte und Kompetenzen, die es zukünftig zu erlernen gilt, als auch die Prüfungsformen werden aktuell diskutiert (Bearman und Luckin 2020; Reinmann 2021; Klar und Schleiss 2024; Furze et al. 2024; Corbin et al. 2025). Die sich verändernden Lern- und Prüfungsprozesse können mittels des SAMR-Modells strukturiert werden. So kann das Lernen und Prüfen entweder erweitert (substituiert oder ergänzt) oder transformiert (modifiziert oder neudefiniert) werden (Puentedura 2006): Eine *Erweiterung* liegt vor, wenn die Technologie als direkter Werkzeugersatz genutzt wird; eine *Transformation* liegt vor, wenn der Einsatz von Technologie eine grundlegende Neugestaltung ermöglicht.

Die Neugestaltung von Lernprozessen wird beispielsweise durch Chatbots umgesetzt, etwa indem diese als Experten für fachliches Lernen, als Assistenten für organisatorische und administrative Fragen sowie als Mentoren für das Anlegen und Erweitern metakognitiver Kompetenzen eingesetzt werden (Wollny et al. 2021). Dabei wirken Chatbots vorteilig für Lernende: Sie unterstützen zeit- und ortsunabhängig durch Feedback, Schritt-für-Schritt-Anleitungen und auf die lernende Person abgestimmte Lerngelegenheiten (Labadze et al. 2023). Werden Lernen und Prüfen als voneinander abhängige Prozesse verstanden (Boud 2000; Boud und Soler 2016), könnten Chatbots auch in Prüfungssituationen integriert werden, sodass insbesondere die Interaktion zwischen genKI und lernender Person Aufschluss über weitere Lernpotenziale geben könnte. Mit dem Ziel, einen Chatbot in eine Prüfungssituation zu übertragen, bedarf es der Entwicklung

eines Prüfungsbots und damit verbunden der Definition von Rahmenbedingungen. Vier Rahmenbedingungen benennen Chang et al. (2023), welche im Folgenden erklärt und auf einen Prüfungsbot übertragen werden:

- *Wissensbereich:* Ein Chatbot kann auf breites, unspezifisches Wissen ausgerichtet sein (generisch), mit einem thematischen Schwerpunkt konzipiert werden (offen) oder innerhalb eines klar abgegrenzten Wissensgebiets arbeiten (geschlossen). Letztgenanntes wäre für einen Prüfungsbot dienlich.
- *Menschliche Unterstützung:* Ein Chatbot kann Anfragen gänzlich ohne menschliche Unterstützung bearbeiten oder eine Weiterleitung an einen Menschen initiieren. Die Funktionsweise eines Prüfungsbots wäre dann abhängig von der Gestaltung des Lern- und Prüfungsprozesses: Ist das Lernen und Prüfen asynchron organisiert, sollte der Prüfungsbot autonom handeln können; findet das Lernen und Prüfen synchron statt, kann eine Weiterleitung an einen Menschen eingerichtet werden.
- *Generierungsmethode der Antwort:* Die Antwort eines Chatbots kann generativ, abruf- oder regelbasiert sein. Alle Generierungsmethoden wären für einen Prüfungsbot vorstellbar; eine generative Erstellung ermöglicht im Vergleich zu den anderen Generierungsmethoden das höchste Mass an Personalisierung.
- *Kommunikationskanal:* Ein Chatbot kann in Form von Text, Sprache oder Bild antworten. Der bevorzugte Kommunikationskanal eines Prüfungsbots wäre abhängig vom In- und Output der Prüfung (Chang et al. 2023, 47–50).

Ein Prüfungsbot kann als Beispiel für die Transformation schriftlicher Prüfungsformen verstanden werden und eine Personalisierung prüfungsbezogener Prozesse ermöglichen. Die damit verbundene (Neu-)Aushandlung von Standardisierung und Personalisierung innerhalb eines Prüfungsprozesses gilt es zu reflektieren.

2.2 Personalisierung von Prüfungsformen

Standardisierung und Personalisierung sind zwei Konzepte, die, so Buzick et al. (2023), an den jeweiligen Enden eines Kontinuums stehen: Während Standardisierung bedeutet, dass alle Lernenden dieselbe Prüfung unter denselben Bedingungen absolvieren mit dem Ziel, die Vergleichbarkeit von Leistungen sicherzustellen, meint Personalisierung, dass die Prüfungssituation an die lernende Person angepasst wird, sodass ein Rückschluss auf die individuelle Kompetenz des Einzelnen möglich wird. Standardisierung schafft *gleiche* und Personalisierung *gerechte* Prüfungsbedingungen (Sireci 2020; Buzick et al. 2023; Walker et al. 2023; Bennett 2024; Nieminen et al. 2025).

Ein Vorschlag, das Verhältnis zwischen Standardisierung und Personalisierung in Prüfungskontexten auszubalancieren, wurde von Saxena et al. (2023) formuliert. In Abhängigkeit von der Funktion der Prüfung soll standardisiert oder personalisiert geprüft werden: Wird der Lernoutput überprüft, müsse standardisiert geprüft werden;

wird der Lernprozess überprüft, könne personalisiert geprüft werden. Eine weitere Herangehensweise an das Spannungsverhältnis kann die Art des Personalisierungsgrades darstellen: Im Sinne der Erweiterung einer standardisierten Prüfungsform kann eine Prüfung kontextualisiert werden (Arslan et al. 2024). Beispielsweise müssen in der medizinischen Ausbildung (internationale) Standards garantiert werden; jedoch differieren die Lehr-Lern-Prozesse z. B. aufgrund der technischen Ausstattung (Bates et al. 2019). Infolgedessen trägt die Kontextualisierung dazu bei, dass die Dichotomie zwischen Standardisierung und Personalisierung aufgelöst und beide Konzepte als komplementär verstanden werden. Für Prüfungen könnte dies bedeuten, dass sowohl Standards gewahrt als auch die kontextualisierte Diversität integriert werden könnte (ebd.).

Der komplementäre Einsatz einer kontextualisierten Prüfung könnte die schriftliche Prüfungsform erweitern. Inwiefern ein solcher Einsatz prüfungsdidaktisch legitim ist, hängt von der Validität einer solchen Prüfung ab.

2.3 Validität in (personalisierten) Prüfungsformen

«Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests» (American Psychological Association et al. 1999, 11). Gemäss Kaldaras et al. (2024) gibt es aktuell kein Rahmenwerk oder keine verbindliche Vorschrift, das oder die zur Sicherstellung und Überprüfung der Validität (gen) KI-basierter Prüfungen und ihrer Ergebnisse genutzt werden könnte. Die Aushandlung von Validität wird so zu einem iterativen und kontextabhängigen Prozess (Kaldaras et al. 2024), wodurch eine nachweisgestützte Abwägung von validitätsfördernden oder -einschränkenden Argumenten erforderlich wird (Dawson et al. 2024). Eine Methode, diesen Aushandlungsprozess nachweisgestützt zu strukturieren, ist mittels einer argumentativen Evidenzkette möglich, welche aus vier Elementen besteht (St-Onge et al. 2017; Huggins-Manley et al. 2022):

- Aussage, eine zu überprüfende Behauptung der Autor:innen;
- Begründung, welche die Aussage argumentativ stützt;
- Annahme, welche die Begründung operationalisierbar und überprüfbar macht;
- Beleg, welcher die Annahme empirisch begründet.

Mit dem Ziel, die Validität personalisierter Prüfungsformen nachweislich zu erschliessen, kann die argumentative Evidenzkette angewendet werden. Hierbei können die theoriegeleiteten Elemente *Aussage* und *Begründung* anhand der Überlegungen von Crooks et al. (1996) und Sinharay et al. (2025) ausgefüllt werden. Eine Prüfung wird dabei folgendermassen definiert:

> «Assessment is depicted as divided into eight conceptually distinct stages, with validation then based on careful scrutiny of each of these stages. The eight stages are likened to eight links of a chain, with weakness of any one link weakening

> the chain as a whole. Further guidance is offered to the validator through identification of several possible validity threats associated with each link.» (Crooks et al. 1996, 266)

Anhand der Definition wird deutlich, dass acht verschiedene Links (administration, scoring, aggregation, generalization, extrapolation, evaluation, decision und impact) trennscharf untersucht werden sollten, um die Validität einer Prüfung sachgemäss beurteilen zu können. Dabei wird jeder Link durch Validity Threats untersuchbar. Im Hinblick auf Prüfungen, die durch die Einflussgrösse der Personalisierung verändert werden, könnten gemäss Sinharay et al. (2025) fünf der acht Links validitätssensitiv sein; demnach gilt es diese zu untersuchen. Hierfür formulierte das Autor:innenteam die Aussage, dass die Personalisierung die Validität entweder beeinträchtigen (vgl. Tabelle 1, Aussage 1) oder aufwerten (vgl. Tabelle 1, Aussage 2) würde. Die Beeinträchtigung würde über den Link scoring feststellbar; die Links administration, evaluation, decision und impact könnten die Validität verbessern. Nachfolgend werden die Begründungen einzeln erklärt:

	Theoriegeleitete Elemente der argumentativen Evidenzkette (St-Onge et al. 2017; Huggins-Manley et al. 2022)	
	Aussage	**Begründung**
Validität einer personalisierten Prüfungsform (Sinharay et al. 2025)	1: Personalisierte Prüfungen haben negative Auswirkungen auf die Prüfungssituation.	*Scoring: questionable score comparability:* Standardisierte Prüfungsaufgaben sind valider als personalisierte Prüfungsaufgaben, weil Leistungen so verglichen werden können.
	2: Personalisierte Prüfungen haben positive Auswirkungen auf die Prüfungssituation.	*Administration: low motivation:* Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende motivierter sind, diese zu lösen.
		Administration: task or response not communicated: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende die zu erledigende Aufgabe besser verstehen.

Tab. 1: Auszug der argumentativen Evidenzkette mit theoriegeleiteten Aussagen und Begründungen zur Validität einer personalisierten Prüfungsform.

Scoring beeinträchtige die Validität einer personalisierten Prüfung, weil durch die Berücksichtigung von Merkmalen der zu prüfenden Person die Vergleichbarkeit zwischen den Lernenden und ihren Kompetenzprofilen nicht mehr gewährleistet werden könne. Im Hinblick auf die Selektions- und Allokationsfunktion sowie die (berufs-)biografische Relevanz von Abschlussprüfungen sei dies problematisch (Sinharay et al. 2025). Der soeben beschriebene Validity Threat «questionable score comparability» des Links scoring kann in der argumentativen Evidenzkette als Begründung für Aussage 1 – Beeinträchtigung der Validität einer Prüfung durch Personalisierung – genutzt werden (vgl. Tabelle 1).

Administration würde die Validität einer personalisierten Prüfung verbessern. Unter diesem Link wird die Durchführung von Prüfungen verstanden und davon ausgegangen, dass insbesondere die Motivation der lernenden Person durch personalisierte Prüfungen steige. Indem Aufgaben an die individuellen Interessen und Kompetenzprofile der lernenden Person angepasst werden, werde die Prüfungsaufgabe als relevanter wahrgenommen und bewusster bearbeitet (Validity Threat: low motivation). Zudem verbessere Personalisierung das Verständnis für die Aufgabenstellung bei einer lernenden Person, da sie diese in ihrer bevorzugten Vermittlung – z. B. schriftlich oder mündlich – erhalten könne. So würden Missverständnisse über die zu bearbeitende Aufgabe reduziert (Validity Threat: task or response not communicated) (Sinharay et al. 2025). Für die in Tabelle 1 dargestellte argumentative Evidenzkette bedeutet dies, dass die Validity Threats «low motivation» und «task or response not communicated» des Links administration als Begründungen für Aussage 2 – Steigerung der Validität einer Prüfung durch Personalisierung – genutzt werden.

Der Link evaluation meint den Umgang mit dem Prüfungsergebnis und wird als validitätssteigernd charakterisiert. So könne eine Person das Ergebnis einer personalisierten Prüfung einfacher nachvollziehen (Validity Threat: poor grasp of assessment information and its limitations) und lediglich die Kompetenzen als (nicht) vorhanden identifizieren, die mit der personalisierten Prüfung abgeprüft wurden. Die personalisierte Prüfungsaufgabe und das dazugehörige Ergebnis würden nicht dazu verleiten, individuelle Leistungen vorschnell auf umfassendere oder nicht intendierte Kompetenzbereiche zu verallgemeinern (Validity Threat: inadequately supported construct interpretation). Zudem würden voreingenommene Interpretationen reduziert (Validity Threat: biased interpretation or explanation) (Sinharay et al. 2025).

Mit dem Link decision wird die Interpretation des Prüfungsergebnisses untersucht. Die Personalisierung von Prüfungen ermögliche eine auf die lernende Person abgestimmte Rückmeldung, mit welcher eine Lehrperson fundierte Entscheidungen über die Lernbegleitung des Einzelnen treffen könne (Validity Threat: poor pedagogical decisions). So werte eine personalisierte Prüfung im Link decision die Validität auf (ebd.).

Eine weitere Steigerung der Validität wird mit impact erkennbar, der als weiterer Link die Auswirkungen der Prüfung auf die lernende Person umfasst. Personalisierte Prüfungen würden individuelle Rückmeldungen ermöglichen, so die Motivation der lernenden Person steigern und von ihr eher als relevant wahrgenommen (Validity Threat: positive consequences not achieved). Zudem würden personalisierte Prüfungen von der lernenden Person eher als fair empfunden, weil sie ihr aktuelles Kompetenzprofil zeigen könnte (Validity Threat: serious negative impact occurs) (ebd.).

Fairness wird als Aspekt im Link impact benannt, kann als Teilmenge von Validität interpretiert werden (Huggins-Manley et al. 2022) und ermöglicht besonders gut, das dichotome (oder komplementäre) Verhältnis von gleichen und gerechten Prüfungsbedingungen zu untersuchen, da der Begriff selbst aus zwei konträren Perspektiven interpretiert und in der Unterrichtspraxis gelebt wird (Murillo und Hidalgo 2020): Wird eine Prüfung für die Evaluierung von Lernprozessen genutzt, wird Fairness im Sinne von gerechten Prüfungsbedingungen verstanden. Wird eine Prüfung für die Evaluierung von Lernoutputs genutzt, wird Fairness im Sinne von gleichen Prüfungsbedingungen verstanden (Tierney 2014). Ergänzend meinen Baniasadi et al. (2023), dass gleiche Prüfungsbedingungen hergestellt werden, indem z. B. formativ geprüft wird und die Bedürfnisse des Einzelnen im Prüfungsprozess beachtet werden (vgl. equitable fairness, Baniasadi et al. 2023). Sind Prüfungsbedingungen gerecht, bedarf es transparent kommunizierter Bewertungskriterien, identischer Prüfungsbedingungen für alle Lernenden und eines objektiven Bewertungsprozesses (vgl. egalitarian fairness, ebd.).

2.4 Forschungsfragen

Basierend auf den vorangegangenen Forschungsperspektiven kann genKI Prüfungen personalisieren und somit schriftliche Prüfungsformen erweitern oder transformieren. Mit dem Ziel, die Wirkung dieser Möglichkeiten zu erschliessen, bedarf es der Untersuchung, inwiefern ein erweiternder (F1) oder transformierender (F2) Personalisierungsgrad das Gütekriterium der Validität einer textbasierten Prüfung gefährdet oder aufwertet.

F1: Inwiefern halten Expert:innen es für valide, eine Prüfung, kontextualisiert mittels genKI, komplementär zur standardisierten Prüfung einzusetzen?

F2: Inwiefern halten Expert:innen es für valide, den Lernenden während der Bearbeitungszeit der Prüfung einen Prüfungsbot zur Verfügung zu stellen?

Zur nachweisgestützten Abwägung von Validität werden die theoriegestützten Elemente der argumentativen Evidenzkette, dargestellt in Tabelle 1, empirisch überprüft.

3. Methodisches Vorgehen

3.1 Untersuchungsgegenstände

Die Untersuchungsgegenstände der Studie waren die standardisierte Prüfung (SP), die kontextualisierte Prüfung (PG A) und der Prüfungsbot (PG B), die von der velpTEC GmbH, einem Weiterbildungsinstitut, zur Verfügung gestellt wurden. Dabei galt:

- SP (vgl. Anhang 1) wurde von einem Experten der velpTEC GmbH für den zweiwöchigen Kurs «KI-Management» unabhängig von der Studie erstellt und prüfte seitdem die Lerninhalte der ersten Lernwoche ab. Über 150 Lernende haben die Prüfung bis September 2025 absolviert.
- PG A (vgl. Anhang 2) wurde, unterstützt durch einen Prompt, mittels ChatGPT-4o erstellt. Als Ausgangslage galt SP. Der Personalisierungsgrad der textbasierten Prüfung bezog sich auf die veränderte Kontextualisierung. So wurde in Abhängigkeit vom berufsbiografischen Hintergrund der lernenden Person ein neues, anwendungsbezogenes und kontextualisiertes Szenario für die Prüfung erstellt. Dieser Personalisierungsgrad wirkt erweiternd (Puentedura 2006).
- PG B gab es in zwei Versionen: Version 1 des PG B lag SP zugrunde, während Version 2 sich auf PG A bezog. So unterschieden sich die Versionen in den unterschiedlichen Prüfungen, auf welche PG B zugriff; im Funktionsumfang stimmten sie aber überein. Das bedeutet, dass die lernende Person während der Lösungserstellung PG B nach Unterstützung fragen konnte. Die konkreten Anfragemöglichkeiten unterschieden sich nach dem Zeitpunkt der Beanspruchung: *Während der Bearbeitung* konnte PG B genutzt werden, um beim Verstehen der Aufgabenstellung, beim Erinnern an vermittelte Lerninhalte, beim Bewusstwerden der Lern- und Kompetenzziele sowie beim Verstehen der Bewertungskriterien zu unterstützen. *Nach der Bearbeitung* konnte PG B genutzt werden mit dem Ziel, ein erstes KI-generiertes Feedback zur eigenen Lösung zu erhalten. Somit agierte PG B innerhalb des Prüfungsumfangs des Kurses «KI-Management» (Wissensbereich), handelte autonom (menschliche Unterstützung), gab generierte Antworten (Generierungsmethode der Antwort) und antwortete textbasiert (Kommunikationskanal) (Chang et al. 2023, 47–50). Die Anfragen waren unlimitiert, wurden gespeichert und waren einsehbar. Dieser Personalisierungsgrad wirkt transformierend (Puentedura 2006).

Die technische Infrastruktur, z. B. Prompt und PG B, war Eigentum der velpTEC GmbH und stand zur Veröffentlichung nicht zur Verfügung. In beiden Fällen wurde ChatGPT-4o als Sprachmodell genutzt.

3.2 Durchführung

Die Durchführung der Studie war in vier Schritte gegliedert: (1) Mithilfe einer Internetrecherche wurden deutschsprachige Professor:innen, Postdocs oder Nachwuchswissenschaftler:innen identifiziert, die mindestens einen fachlichen Schwerpunkt in «Angewandter Künstliche Intelligenz», «Künstlicher Intelligenz», «Fachdidaktik für Informatik», «Bildungstechnologie», «E-Assessment», «Mensch-KI-Kollaboration» oder «KI und Erwachsenenbildung» aufwiesen. (2) Erfüllte eine Person dieses Kriterium, wurde sie per E-Mail eingeladen. Nahm sie die Einladung an, erhielt die Person sowohl den digitalen Zugang zur Studie als auch die ausführlichen Studieninformationen sowie datenschutzrechtlichen Hinweise. (3) Zur Beantwortung von F1 – Vergleich von standardisierter und kontextualisierter textbasierter Prüfung – wurde ein Kontrollgruppendesign entwickelt. So wurde die teilnehmende Person einer von zwei Gruppen (SP oder PG A) zugeordnet und erhielt daraufhin Zugang zu einer der zwei Webseiten, die in Aufbau und Funktionalität identisch waren und sich lediglich in der zur Verfügung gestellten Prüfung und dem dazugehörigen Bewertungsbogen unterschieden.[1] Darüber hinaus enthielten beide Webseiten eine persönliche Anrede, Instruktionen und eine individuelle Studien-ID[2] sowie Informationen zum Kurs und den Lernzielen sowie den Link zum Fragebogen für F1.[3] Der Fragebogen für F1 war sowohl für Webseite 1 als auch 2 derselbe, um ein Kontrollgruppendesign umzusetzen. Die Fragebogenitems wurden mithilfe einer 5er Likert-Skala von der teilnehmenden Person beantwortet. Bei der Abstimmung galt: 1 bedeutete «Ich stimme überhaupt nicht zu» und 5 «Ich stimme voll und ganz zu». (4) Mit dem Ziel, den Einsatz von PG B zu untersuchen (vgl. F2), erhielten die teilnehmenden Personen Zugang zu PG B[4] sowie zum Fragebogen für F2.[5] Der Fragebogen für F2 war für beide Webseiten identisch. Sowohl die Fragebogenitems als auch die Nutzung einer 5er Likert-Skala pro Fragebogenitem waren nahezu identisch mit F1. Die Fragebogenitems unterschieden sich von F1 in ihren Formulierungen. So wurde speziell auf die Nutzung von PG B eingegangen. Zugleich wurden Vorab-Fragen zur Nutzungshäufigkeit und -intensivität gestellt sowie ein Fragebogenitem zur *interactional fairness* ergänzt, da PG B mit der lernenden Person interagiert.

1 Eine teilnehmende Person mit Zugang zu Webseite 1 erhielt Einsicht in SP, während eine teilnehmende Person mit Zugang zu Website 2 Einsicht in PG A erhielt.

2 Die Studien-ID war eine zwölfstellige Abfolge von Buchstaben und Zahlen. Es war kein Rückschluss von der teilnehmenden Person auf die Studien-ID möglich.

3 Zu Beginn des Fragebogens wurde die Studien-ID abgefragt, um die mehrfache Einreichung von Antworten zu verhindern.

4 PG B war in seinen Funktionalitäten für beide Webseiten identisch. Eine teilnehmende Person, die im Durchführungsschritt (3) Zugang zu SP hatte, erhielt Zugriff auf PG B, der sich auf SP bezog. Eine teilnehmende Person, die im Durchführungsschritt (3) Zugang zu PG A hatte, erhielt Zugriff auf PG B, der sich auf PG A bezog.

5 Auch an dieser Stelle war die Eingabe der Studien-ID erforderlich, um ein einmaliges Antwortverhalten pro teilnehmender Person zu garantieren.

3.3 Stichprobe

Zwischen Juli und September 2025 wurden insgesamt 105 Expert:innen per Mail angeschrieben. Davon gaben 32 eine positive Rückmeldung und 18 füllten die digitalen Fragebögen vollständig aus, wobei 9 Expert:innen SP und 9 Expert:innen PG A bearbeiteten. Rückmeldungen zu den Fragebögen waren bis zum 26. 09. 2025 möglich.

3.4 Auswertungsdesign für F1

Mit dem Ziel, F1 zu beantworten, wird die in Tabelle 1 dargestellte argumentative Evidenzkette um die Spalten *Annahme* und *Beleg* erweitert (vgl. Tabelle 2): Grundlegend gilt es, die von Sinharay et al. (2025) getroffenen *Aussagen* und *Begründungen* zur Validität personalisierter Prüfungen mithilfe eines Kontrollgruppendesigns zu überprüfen. So werden in der Spalte *Annahme* die entwickelten Fragebogenitems hinterlegt, die Rückschluss auf die *Begründung* und somit auf die *Aussage* geben sollen. Das Antwortverhalten der teilnehmenden Personen und die dadurch möglichen statistischen Auswertungen werden in der Spalte *Beleg* erfasst (vgl. Tabelle 2).

Konkret wird zur Überprüfung der Unterschiede zwischen SP und PG A eine quantitative Vergleichsanalyse mit unabhängigen Stichproben durchgeführt. Hierfür werden die Ergebnisse nach Hedges' g berechnet (Hedges und Olkin 1985). Dadurch kann die Stärke und Richtung der Unterschiede unabhängig von der Stichprobengrösse beurteilt werden. Die Interpretation der Effektstärken folgt der Klassifikation nach Cohen (1988): klein: $g \approx 0{,}2$; mittel: $g \approx 0{,}5$; gross: $g \geq 0{,}8$. Zur inferenzstatistischen Abklärung wird aufgrund der geringen Stichprobengrösse der Wilcoxon-Rangsummentest für unabhängige Stichproben verwendet (Wilcoxon 1945; Mann und Whitney 1947). Das Signifikanzniveau wird mit $p < ,05$ angesetzt (vgl. Tabelle 2).

Theoriegeleitet		Empirische Studie: Auswertungsdesign für F1						
Aussage	Begründung	Annahme *(Fragebogenitems)*	Beleg *(Studienergebnisse)*					
			Gruppe	MW	SD	g	W	p
1. Personalisierte Prüfungen haben negative Auswirkungen auf die Prüfungssituation.	*Scoring: questionable score comparability:* Standardisierte Prüfungsaufgaben sind valider als personalisierte Prüfungsaufgaben, weil Leistungen so verglichen werden können.	Die Prüfungsaufgabe kann mithilfe des Bewertungsbogens bewertet werden.	SP					
			PG A					
		Der Bewertungsbogen ermöglicht eine standardisierte Auswertung der Prüfungsaufgabe.	SP					
			PG A					

Tab. 2: Auszug der argumentativen Evidenzkette für F1.

Die Kombination aus Wilcoxon-Rangsummentest und Effektstärkenanalyse ermöglicht sowohl eine statistische Prüfung auf Unterschiede als auch eine inhaltliche Einordnung der Befunde im Hinblick auf die theoretischen Validitätsannahmen. Aufgrund der begrenzten Stichprobengrösse und des explorativen Studiendesigns erfolgt die Interpretation primär richtungs- und theoriebasiert (Lakens 2013). Hierbei hat der Signifikanztest einen rein explorativen Charakter. Demnach kann auch ein nicht signifikanter Effekt nach seiner Effektgrösse bewertet werden.

3.5 Auswertungsdesign für F2

Inwiefern der Einsatz eines Prüfungsbots valide ist (F2), wird mit einer weiteren argumentativen Evidenzkette erschlossen. Da die Funktionalitäten des Prüfungsbots für beide Kontrollgruppen identisch waren, liegt für F2 kein Kontrollgruppendesign vor und alle teilnehmenden Personen werden als Ganzes betrachtet. Die zugrundeliegende Prüfung (SP oder PG A) des Prüfungsbots spielt keine Rolle. Die Berechnungen erfolgen auf Basis eines Inter-Item-Vergleichs. Für jedes Fragebogenitem werden Mittelwert (MW) und Standardabweichung (SD) berechnet, um zentrale Tendenzen und die Streuung der Einschätzungen zu beschreiben. Die MW zeigen das allgemeine Zustimmungsausmass auf, während die SD Hinweise auf die Homogenität oder Heterogenität der Urteile geben (Bortz und Döring 2006). Die Ergebnisse werden deskriptiv-explorativ interpretiert, um wahrnehmbare Bewertungsmuster und Spannungsfelder zwischen Standardisierung und Personalisierung zu identifizieren.

4. Ergebnisse

4.1 F1: Erweiterter Personalisierungsgrad

Die erste Forschungsfrage untersucht, inwiefern die Kontextualisierung der Prüfungsaufgabe die Validität einer Prüfung beeinflusst. In Anhang 3 sind neben der vollständigen argumentativen Evidenzkette inkl. MW, SD und Effektstärken (Hedges' g) die Ergebnisse des Wilcoxon-Rangsummen-Tests für F1 dargestellt.

In Anhang 3 zeigen die Ergebnisse über alle Fragebogenitems hinweg eine hohe Zustimmung zu den Aussagen (MW ≈ 4,0). Damit werden die Prüfungsaufgaben in beiden Varianten insgesamt positiv bewertet. Hohe MW – sowohl in SP (MW = 4,00 bis 4,56) als auch PG A (MW = 4,00 bis 4,33) – finden sich in Item 1, 4, 5, 6, 8 und 15 bis 18; schwerpunktmässig sind die Items den Links administration und impact zuordenbar. Vergleichsweise geringere MW in beiden Prüfungsaufgaben zeigen sich in Item 7 (MW SP = 3,89; PG A = 3,89), 9 (MW SP = 3,56; PG A = 2,89), 13 (MW SP = 3,89; PG A = 3,67), 14 (MW SP = 2,67; PG A = 2,67), 19 (MW SP = 3,89; PG A = 3,33) und 20 (MW SP = 3,33; PG A = 3,11). Sie sind grösstenteils den Links evaluation und impact zuordenbar.

Die Ergebnisse der Effektstärken nach Hedges' g ermöglichen einen differenzierten Vergleich. Dabei repräsentieren negative g-Werte eine höhere Zustimmung für die kontextualisierte Prüfungsaufgabe (PG A), positive Werte für die standardisierte Variante (SP). Die Items 1, 7, 14 und 15 weisen g = 0,00 auf. Entsprechend lassen sich zwischen SP und PG A keine Unterschiede in Bezug auf die Bewertbarkeit (Item 1), diagnostische Rückschlüsse (Item 7), die Empfehlungsrelevanz (Item 14) sowie die Fairness (Item 15) identifizieren. In elf der 20 Items finden sich geringe Unterschiede zwischen SP und PG A ($g < 0{,}40$). Der Schwerpunkt der Richtungseffekte liegt dabei häufiger zugunsten von SP. Drei der 20 Items (Item 3, 18 und 19) weissen eine mittlere und zwei (Item 9 und 11) eine grosse Effektgrösse auf. Item 3 ($g = -0{,}57$) und Item 11 ($g = -1{,}36$) richten sich zugunsten von PG A und ermöglichen eine Aussage über motivationale und authentizitätsstiftende Merkmale von kontextualisierten Prüfungen, wobei ausschliesslich die Effektstärke von Item 11 signifikant ist. Item 9 ($g = 0{,}84$), 18 ($g = 0{,}56$) und 19 ($g = 0{,}64$) weisen auf SP gerichtete positive Effektstärken auf. Die Items beziehen sich auf den Aspekt der ausschliesslichen Kompetenzmessung und auf die Abfrage zur wahrgenommenen Fairness.

4.2 F2: Transformativer Personalisierungsgrad

Die zweite Forschungsfrage untersucht, inwiefern der Einsatz eines Prüfungsbots die Validität einer Prüfung gefährden oder aufwerten kann (vgl. Anhang 4 und 5). Der Ansatz ist explorativ. Insgesamt probierten 17 der 18 Expert:innen den Prüfungsbot in unterschiedlichen Intensitäten aus und hatten damit Zugriff auf den Fragebogen, der in Anhang 5, Spalte Annahme, einsehbar ist.

Für eine strukturierte Analyse müssen die Fragebogenitems 7, 16 bis 20 vorab differenziert betrachtet werden: Gemäss der entwickelten argumentativen Evidenzkette in Anhang 5 belegen diese Items Aussage 2, dass eine personalisierte Prüfung positive Auswirkungen auf die Prüfungssituation hat. Aufgrund der inhaltlichen Formulierung, z. B. fehlerhafte Informationen, inhaltliche Vorteile sowie Ablenkung von prüfungsrelevanten Informationen, wird jedoch die Gefährdung von Validität impliziert und quantifiziert. Somit werden im Folgenden die Items 7 und 16 bis 20 Aussage 1 – personalisierte Prüfungen haben negative Auswirkungen auf die Prüfungssituation – zugeordnet.

Bei den Items, welche die Gefährdung von Validität quantifizieren, liegt der kleinste MW bei Item 1 mit 2,35 (Gefährdung der standardisierten Auswertung) und der grösste MW bei Item 17 mit 3,88 (Änderung des Anforderungsniveaus) vor. Generell sind die Streuungen über all diese Items hinweg mittel bis hoch – die höchste Streuung liegt bei Item 19 mit 1,50 (inhaltlicher Vorteil) und die geringste Streuung bei Item 17 mit 1,05 (Änderung des Anforderungsniveaus). Hervorzuheben sind die Items zur Bewertung (1, 2 und 3) sowie zur Fairness (18 bis 20), die sehr hohen Streuungen unterliegen (vgl. Anhang 5).

Bei den zu bewertenden Begründungen, die eine Aufwertung der Validität durch den Prüfungsbot implizieren, hat mindestens ein Item aus jedem Link (administration, evaluation, decision, impact) einen MW von > 4,00: Für administration ist es Item 6 (Unterstützung beim Verstehen der Aufgabe), für evaluation ist es Item 10 (Planung des weiteren Lernwegs), für decision ist es Item 12 (nützliches Feedback während der Prüfung) und für impact ist es Item 21 (diskriminierungsfreie Kommunikation) (vgl. Anhang 5). Auffällig ist, dass bei den Items mit den höchsten MW vergleichsweise niedrige bis mittlere Streuungen vorliegen (Item 6: MW = 4,18, SD = 1,02; Item 10: MW = 4,12, SD = ,86; Item 12: MW = 4,06, SD = ,90; Item 21: MW = 4,06, SD = ,66).

Zuletzt zeigt der Vergleich zu F1, dass die Bewertungen bei F2 deutlich stärker streuen. Während die Kontextualisierung (F1) überwiegend konsistente, leicht positive Einschätzungen erhielt, zeigen die Items zur Nutzung des Prüfungsbots (F2) grössere Divergenzen in den Urteilen der Expert:innen.

5. Diskussion

5.1 F1: Erweiterter Personalisierungsgrad

Die Diskussion der Ergebnisse erfolgt unter der besonderen Beachtung der Limitationen der Studie, die ausführlich in Kapitel 5.3 erläutert werden. Das Ziel, die Validität personalisierter Prüfungsformen zu erfassen, erfolgte explorativ: sowohl Validität argumentativ zu erfassen und so das Gütekriterium innerhalb eines iterativen Aushandlungsprozesses zu untersuchen (Kaldaras et al. 2024; Dawson et al. 2024), als auch Prüfungsformen zu

personalisieren, stellen neue Möglichkeiten dar, die den Bedarf weiterer theoriegeleiteter und empirischer Forschungen im Hinblick auf die Themen Validität und Personalisierung von Prüfungsformen aufzeigen.

Für den erweiterten Personalisierungsgrad kann Aussage 1 zurückgewiesen werden, wonach personalisierte Prüfungen negative Auswirkungen auf die Prüfungssituation haben und so die Validität gefährden (Sinharay et al. 2025): Sowohl die geringen Effektgrössen des Links scoring als auch Item 16 (Überprüfung des Lern- und Kompetenzziels) zeigen, dass eine standardisierte Bewertung beider Prüfungen (SP und PG A) möglich ist und der Einsatz eines kontextualisierten Rahmentexts die Validität der Prüfung nicht beeinträchtigt. Somit ist ein komplementärer Einsatz möglich. Mit dem Ziel, die Vorgehensweise der velpTEC GmbH zu verallgemeinern und so in unterschiedlichen Lernszenarios an einen individualisierten Lernprozess einen individualisierten Prüfungsprozess zu knüpfen, kann folgendes Vorgehen abgeleitet werden: Die von einer Lehrkraft entwickelte standardisierte Prüfungsaufgabe könnte auf unveränderbare und veränderbare Elemente untersucht werden. Unveränderbar sollten dabei die Aufgaben- und Bearbeitungslogik, das Anforderungsniveau und die Bewertungskriterien sein. Veränderbar kann das Szenario bzw. der Rahmentext der Prüfung sein. Mittels eines Prompts, welcher bspw. die Interessen der lernenden Person berücksichtigt und sich an SP orientiert, könnte zukünftig durch genKI eine kontextualisierte Prüfung entwickelt werden.

Dass diese Erweiterung lerndienlich sein kann, wird durch die mittlere Effektgrösse von Item 3 (Motivation) und die hohe, signifikante Effektgrösse von Item 11 (Authentizität) gezeigt. Mit dem Ziel, diese Effekte zu nutzen, scheint der Einsatz einer kontextualisierten Prüfungsform unabdingbar in Lerneinheiten, in welchen sowohl die beruflichen Hintergründe der Lernenden heterogen sind als auch das Lernen standardisiert ist und so die Relevanz der Lerninhalte von den Lernenden hinterfragt wird. Erfahrungsgemäss finden sich diese Bedingungen in unternehmensinternen Weiterbildungen zu verpflichtenden und teamunabhängigen Weiterbildungsinhalten, z. B. Datenschutz oder KI-Management. Mit dem Einsatz kontextualisierter Prüfungen in diesen Weiterbildungen könnte bspw. im Sinne des «assessment for learning»-Ansatzes (Boud 2000; Boud und Soler 2016) die kontextualisierte Prüfung den Lernprozess unterstützen und diesen in besonderem Mass durch lernförderliche Effekte wie Motivation und Authentizität aufwerten.

Aussage 2, wonach personalisierte Prüfungen positive Auswirkungen auf die Prüfungssituation haben und demnach die Validität einer Prüfung aufwerten, kann für den erweiterten Personalisierungsgrad nicht eindeutig bestätigt werden. Die von Sinharay et al. (2025) angenommene Dominanz personalisierter Prüfungen gegenüber standardisierten Prüfungen konnte nicht in den Links administration, evaluation, decision und impact festgestellt werden. Aufgrund der häufig positiven Effektgrössen könnte sogar SP als valider im Vergleich zu PG A charakterisiert werden. Da aber die Mittelwerte ähnlich

hoch und die Effektgrössen meist gering sind, ist der gewählte Personalisierungsgrad möglicherweise keine kritische Einflussgrösse für die Validität der Prüfung.[6] Infolgedessen könnte im Sinne eines selbstverantwortlichen Lehr-Lern-Prozesses die Entscheidung der erwachsenen lernenden Person überlassen werden, inwiefern sie eine standardisierte oder von genKI entwickelte kontextualisierte Prüfung bearbeiten möchte. Indem die Entscheidung auf die lernende Person übertragen wird, kann auch den diffusen Ergebnissen der Fairness-Items (15 bis 20) Rechnung getragen werden: Interessanterweise wird das erste Item zur Abfrage von Fairness in beiden Varianten gleich eingeschätzt; die Prüfung sei grundsätzlich fair (MW SP = 4,00; PG A = 4,00). Jedoch zeigen die fünf Items zur Abfrage von gleichen Prüfungsbedingungen ausnahmslos positive Effektgrössen, sodass gemäss den Expert:innen die kontextualisierte Prüfung eher zu ungleichen Prüfungsbedingungen beitragen würde. Da der erste Impuls zur Fairness (Item 15) und lediglich zwei der fünf Items (Item 18 und 19) mittelstarke positive Effektgrössen aufweisen, scheint der Einsatz dennoch legitim, sodass die lernende Person mit der Entscheidung der Prüfungsaufgabe selbst abwägen kann, welches Verständnis von Fairness sie innerhalb der Prüfung umsetzen möchte.

Abschliessend wird mit der Übertragung der Entscheidung auf die erwachsene lernende Person deutlich, dass im erweiterten Personalisierungsgrad die beiden Konzepte Personalisierung und Standardisierung komplementär zueinander stehen (Bates et al. 2019). Ist die Voraussetzung – eine einheitliche Bewertung durchführen zu können – gegeben, sind beide Prüfungsvarianten grundsätzlich valide und deren Einsatz legitim.

5.2 F2: Transformativer Personalisierungsgrad

Aussage 1, wonach personalisierte Prüfungen negative Auswirkungen auf die Prüfungssituation haben, kann für den transformativen Personalisierungsgrad nicht eindeutig abgelehnt werden (Sinharay et al. 2025): Besonders die hohen SD quantifizieren die Uneinigkeit unter den Expert:innen – z. B. hervorgerufen durch unterschiedliche Nutzungsintensitäten während der Testphase oder fehlende externe Anhaltspunkte im Umgang mit der transformierenden Technologie. Infolgedessen erfordert der Einsatz eines Prüfungsbots die unbedingte Offenlegung der Funktionalitäten für alle Nutzenden (Long und Magerko 2020). Grundsätzlich zeigen die Ergebnisse, dass durch den Prüfungsbot die standardisierte Auswertung der Prüfung (Item 1) möglich ist, aber die Eigenleistung der lernenden Person (Items 3, 17 und 19) infrage gestellt wird. Ohne Einsicht in die Chatprotokolle sind zwei Schlussfolgerungen möglich: (1) Hat der Prüfungsbot ausschliesslich gemäss seinen Handlungsanweisungen unterstützt, ist es möglich, dass die Expert:innen unter Eigenleistung einen Lösungsprozess verstanden haben, der gänzlich

6 Die Interpretation einzelner Items, z. B. Item 7 (Rückschluss auf den aktuellen Lernstand, g = 0.00, MW = 3,89), zeigt möglicherweise, dass nicht der Personalisierungsgrad per se überdacht werden muss, sondern SP als Ausgangsprüfung. Somit bleibt die Lehrkraft zentral in der Erstellung valider Prüfungsformen verankert.

ohne den Einsatz von genKI funktioniert. Aufgrund der Durchdringung von genKI in nahezu allen Lebensbereichen bedarf es gemäss Luo (2024) einer Nuancierung dieses Konzepts von Originalität. Mensch und Maschine werden zusammenarbeiten – auch in Prüfungskontexten. (2) Hat der Prüfungsbot umfassend inhaltlich unterstützt, bedarf es der Anpassung der Funktionslogik des Prüfungsbots. Im Rahmen des Prompts konnte der Prüfungsbot auf alle Arten von inhaltlichen Anfragen reagieren, z. B. «Ich verstehe Aufgabe 3 nicht». Zukünftig könnte überlegt werden, dass der Prüfungsbot ausschliesslich auf aufgabenbezogene Rückfragen der Lernenden reagiert – etwa im Sinne von: «Habe ich richtig verstanden, dass ich in Aufgabe 3 […] tun soll?». Bestätigt der Prüfungsbot die Interpretation, bleibt die Eigenleistung der lernenden Person unberührt; verneint er, kann sie die Aufgabenstellung reflektieren und gegebenenfalls erneut nachfragen. Auf diese Weise würde der Prüfungsbot nicht inhaltlich in den Lösungsprozess eingreifen, sondern ausschliesslich die Verständnissicherung fördern und möglicherweise die Eigenleistung der lernenden Person weniger reduzieren.

Aussage 2, dass personalisierte Prüfungen positive Auswirkungen auf die Prüfungssituation haben, kann grösstenteils zugestimmt werden: Mit der Nutzung des Prüfungsbots können umfangreiche Feedbackprozesse während und nach der Prüfung initiiert werden. Mit der Einsicht in den Chatverlauf können Nachfragen und ggf. Lerndefizite transparent nachvollzogen und daraufhin der weitere Lernweg geplant werden. Da der Chatbot diskriminierungsfrei kommuniziert (Item 21), wird der Prüfungsbot vermeintlich von den Lernenden als niedrigschwelliges und wertfreies Unterstützungsmedium angesehen. So können die Vorteile, die mit dem Einsatz eines Chatbots im Lernprozess einhergehen, auch auf den Prüfungsprozess übertragen werden (Labadze et al. 2023).

Da die Nutzung des Prüfungsbots gänzlich von der lernenden Person abhängig ist und keine Voraussetzung für eine sehr gute Leistungserbringung darstellt, kann das Verhältnis von Standardisierung und Personalisierung als komplementär charakterisiert werden (Bates et al. 2019).

5.3 Limitationen der Studie

Die Untersuchung der argumentativen Validität wurde theoriegeleitet für F1 und F2 hergeleitet. Nichtsdestotrotz sind methodische Ungenauigkeiten nicht auszuschliessen, da die aktuelle Literatur zur argumentativen Validität begrenzt ist und insbesondere die Aushandlung der Inhaltsvalidität mit unterschiedlichen quantifizierenden Berechnungen vorgenommen wird (Santoso et al. 2022).

Für die Personalisierungsgrade gilt, dass die Ergebnisse abhängig vom Prompt und Prüfungsbot der velpTEC GmbH sind. Aufgrund der geringen Anzahl der Expert:innen ist die Aussagekraft der Ergebnisse begrenzt. Es bedarf Anschlussforschungen mit einer umfangreicheren Stichprobe, dem Einbezug der Lernenden als Teilnehmer:innen, da

diese von einem individualisierten Lern- und Prüfungsprozess profitieren sollen (Circi et al. 2023), und qualitativen Analysen mit dem Ziel, unerwartete Themen zu identifizieren und zu adressieren, die mit dem Einsatz einer personalisierten Prüfung in der Unterrichtspraxis einhergehen.

Zuletzt ist zu beachten, dass die untersuchten Prüfungsaufgaben per se von z. B. ChatGPT beantwortet werden können und somit «nicht besonders geeignet für eine leistungsentscheidende Prüfung» (Dawson et al. 2024, 1012) sind.

6. Fazit

Insgesamt untersuchte die Studie die argumentative Validität von zwei verschiedenen Personalisierungsgraden – erweitert (F1) und transformiert (F2) – einer textbasierten Prüfungsform. Hierfür wurden Expert:innen (n = 18) mittels zweier Fragebögen befragt. Die Ergebnisse dienten als empirische Grundlage für die in der Studie erstellte argumentative Evidenzkette, mit welcher das Konzept der Validität überprüft werden konnte.

Die Ergebnisse zeigen, dass beide Personalisierungsgrade die Validität entsprechend der Expert:inneneinschätzung nicht signifikant beeinflussen und somit ein komplementärer Einsatz in Prüfungen möglich wäre. Während eine kontextualisierte Prüfung motivational und authentizitätsstiftend wirken kann, kann der Prüfungsbot ein transparentes Feedback während und nach der Prüfung ermöglichen.

Insgesamt beabsichtigte die Studie, einerseits eine ergänzende Perspektive zur Überarbeitung von schriftlichen Prüfungsformen zu eröffnen mit dem Ziel, personalisierte *und* valide Prüfungen zu entwickeln. Andererseits wurde der Versuch unternommen, Validität argumentativ zu erfassen und diese begründet abzuwägen (Dawson et al. 2024). Hieran anschliessend könnten Forschungen ansetzen, die das Konzept der argumentativen Validität auf die Unterrichtspraxis übertragen und so zur Ausdifferenzierung der diagnostischen Kompetenzen von Lehrkräften beitragen.

Literatur

Ali, Farhan, Doris Choy, Shanti Divaharan, Hui Yong Tay, und Wenli Chen. 2023. «Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering». *Learning: Research and Practice* 9 (2): 135–47. https://doi.org/10.1080/23735082.2023.2258886.

Alier, Marc, Francisco García-Peñalvo, und Jorge D. Camba. 2024. «Generative Artificial Intelligence in Education: From Deceptive to Disruptive». *International Journal of Interactive Multimedia and Artificial Intelligence* 8 (März): 5. https://doi.org/10.9781/ijimai.2024.02.011.

American Psychological Association, National Council on Measurement in Education, und Joint Committee on Standards for Educational and Psychological Testing (U. S.). 1999. *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Arslan, Burcu, Blair Lehman, Caitlin Tenison, Jesse R. Sparks, Alexis A. López, Lin Gu, und Diego Zapata-Rivera. 2024. «Opportunities and challenges of using generative AI to personalize educational assessment». *Frontiers in Artificial Intelligence* 7. https://doi.org/10.3389/frai.2024.1460651.

Baniasadi, Ali, Keyvan Salehi, Ebrahim Khodaie, Khosrow Bagheri Noaparast, und Balal Izanloo. 2023. «Fairness in Classroom Assessment: A Systematic Review». *The Asia-Pacific Education Researcher* 32 (Januar): 91–109. https://doi.org/10.1007/s40299-021-00636-z.

Bates, Joanna, Brett Schrewe, Rachel Ellaway, Pim Teunissen, und Chris Watling. 2019. «Embracing standardisation and contextualisation in medical education». *Medical Education* 53 (1). https://doi.org/10.1111/medu.13740.

Bearman, Margaret, und Rosemary Luckin. 2020. «Preparing University Assessment for a World with AI: Tasks for Human Intelligence». In *Re-imagining University Assessment in a Digital World,* herausgegeben von Margaret Bearman, Phillip Dawson, Rola Ajjawi, Joanna Tai, und David Boud, 49–63. Cham: Springer. https://doi.org/10.1007/978-3-030-41956-1_5.

Bennett, Randy E. 2024. «Personalizing Assessment: Dream or Nightmare?» *Educational Measurement: Issues and Practice* n/a (n/a). https://doi.org/10.1111/emip.12652.

Bortz, Jürgen, und Nicola Döring. 2006. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler (Sonderausgabe der 4. Auflage)*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-33306-7.

Boud, David. 2000. «Sustainable Assessment: Rethinking assessment for the learning society». *Studies in Continuing Education* 22 (2): 151–67. https://doi.org/10.1080/713695728.

Boud, David, und Rebeca Soler. 2016. «Sustainable assessment revisited». *Assessment & Evaluation in Higher Education* 41 (3): 400–13. https://doi.org/10.1080/02602938.2015.1018133.

Buzick, Heather M., Jodi M. Casabianca, und Melissa L. Gholson. 2023. «Personalizing Large-Scale Assessment in Practice». *Educational Measurement: Issues and Practice* 42 (2): 5–11. https://doi.org/10.1111/emip.12551.

Chan, Kuang Wen, Farhan Ali, Joonhyeong Park, Kah Shen Brandon Sham, Erdalyn Yeh Thong Tan, Francis Woon Chien Chong, Kun Qian, und Guan Kheng Sze. 2025. «Automatic item generation in various STEM subjects using large language model prompting». *Computers and Education: Artificial Intelligence* 8 (Juni): 100344. https://doi.org/10.1016/j.caeai.2024.100344.

Chang, Jina, Joonhyeong Park, und Jisun Park. 2023. «Using an Artificial Intelligence Chatbot in Scientific Inquiry: Focusing on a Guided-Inquiry Activity Using Inquirybot». *Asia-Pacific Science Education* (Leiden, Niederlande) 9 (1): 44–74. https://doi.org/10.1163/23641177-bja10062.

Circi, Ruhan, Juanita Hicks, und Emmanuel Sikali. 2023. «Automatic item generation: foundations and machine learning-based approaches for assessments». *Frontiers in Education* 8. https://doi.org/10.3389/feduc.2023.858273.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edition. New York: Routledge.

Corbin, Thomas, Phillip Dawson, und Danny Liu. 2025. «Talk is cheap: why structural assessment changes are needed for a time of GenAI». *Assessment & Evaluation in Higher Education* Mai 15: 1–11. https://doi.org/10.1080/02602938.2025.2503964.

Crooks, Terry, Michael Kane, und Allan Cohen. 1996. «Threats to the Valid Use of Assessments». *Assessment in Education: Principles, Policy & Practice* 3 (November): 265–86. https://doi.org/10.1080/0969594960030302.

Dawson, Phillip, Margaret Bearman, Mollie Dollinger, und David Boud. 2024. «Validity matters more than cheating». *Assessment & Evaluation in Higher Education* 49 (7): 1005–16. https://doi.org/10.1080/02602938.2024.2386662.

Furze, Leon, Mike Perkins, Jasper Roe, und Jason MacVaugh. 2024. «The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment». *Australasian Journal of Educational Technology*, Online-Vorab-Publikation, Oktober 16. https://doi.org/10.14742/ajet.9434.

Hedges, Larry, und Ingram Olkin. 1985. «Statistical Methods in Meta-Analysis». Orlando: Academic Press. https://doi.org/10.1016/C2009-0-03396-0.

Huggins-Manley, A. Corinne, Brandon M. Booth, und Sidney K. D'Mello. 2022. «Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments». *Journal of Educational Measurement* 59 (3): 362–88. https://doi.org/10.1111/jedm.12334.

Kaldaras, Leonora, Hope O. Akaeze, und Mark D. Reckase. 2024. «Developing valid assessments in the era of generative artificial intelligence». *Frontiers in Education* 9. https://doi.org/10.3389/feduc.2024.1399377.

Klar, Maria, und Johannes Schleiss. 2024. «Künstliche Intelligenz im Kontext von Kompetenzen, Prüfungen und Lehr-Lern-Methoden: Alte und neue Gestaltungsfragen». JFMH23: Spannungsfeld Der Digitalen Kompetenz. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 58 (JFMH2023): 41–57. https://doi.org/10.21240/mpaed/58/2024.03.24.X.

Labadze, Lasha, Maya Grigolia, und Lela Machaidze. 2023. «Role of AI chatbots in education: systematic literature review». *International Journal of Educational Technology in Higher Education* 20 (1): 56. https://doi.org/10.1186/s41239-023-00426-1.

Lakens, Daniel. 2013. «Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs». *Frontiers in Psychology* 4. https://doi.org/10.3389/fpsyg.2013.00863.

Long, Duri, und Brian Magerko. 2020. «What is AI Literacy? Competencies and Design Considerations». *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '20, 1–16. https://doi.org/10.1145/3313831.3376727.

Luo (Jess), Jiahui. 2024. «A critical review of GenAI policies in higher education assessment: a call to reconsider the ‹originality› of students' work». *Assessment & Evaluation in Higher Education* 49 (5): 651–64. https://doi.org/10.1080/02602938.2024.2309963.

Mann, Henry, und Donald Whitney. 1947. «On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other». *The Annals of Mathematical Statistics* 18 (1): 50–60. https://doi.org/10.1214/aoms/1177730491.

Murillo, F. Javier, und Nina Hidalgo. 2020. «Fair student assessment: A phenomenographic study on teachers' conceptions». *Studies in Educational Evaluation* 65 (Juni): 100860. https://doi.org/10.1016/j.stueduc.2020.100860.

Nicola-Richmond, Kelli, Phillip Dawson, Helen Partridge, und Susie Macfarlane. 2025. «It takes a village… Program-wide approaches to redesigning assessment in a time of generative artificial intelligence (GenAI)». Curriculum and Assessment Design. *Journal of University Teaching and Learning Practice*. https://doi.org/10.53761/zpp2ja61.

Nieminen, Juuso Henrik, Mollie Dollinger, und Rachel Finneran. 2025. «‹There was very little room for me to be me›: the lived tensions between assessment standardisation and student diversity». *Assessment & Evaluation in Higher Education* 50 (2): 308–22. https://doi.org/10.1080/02602938.2024.2388699.

Puentedura, Ruben. 2006. «Transformation, Technology, and Education». http://hippasus.com/resources/tte/puentedura_tte.pdf.

Reinhold, Lhea, und Marion Händel. 2025. «Bewertung durch eine künstliche Intelligenz? Auswertungs- und Interpretationsobjektivität von ChatGPT-4o bei der Bewertung von Lerntagebucheinträgen». MEDIDA24: 1. Tagungsband des AK Mediendidaktik der DGfE-Sektion Medienpädagogik. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 65 (MEDIDA24): 227–50. https://doi.org/10.21240/mpaed/65/2025.08.03.X.

Reinmann, Gabi. 2021. «Prüfungstypen, -formate, -formen oder -szenarien?» *Impact Free* 36. https://gabi-reinmann.de/wp-content/uploads/2021/06/Impact_Free_36.pdf.

Santoso, Agung, Alice Whita Savira, und Robertus Landung Eko Prihatmoko. 2022. «Validity evidence based on content: Controversies and quantification». 5–16. https://doi.org/10.37517/978-1-74286-697-0-01.

Saxena, Ritcha, Kevin Carnevale, Oleg Yakymovych, Michael Salzle, Kapil Sharma, und Ritwik Raj Saxena. 2023. «Precision, Personalization, and Progress: Traditional and Adaptive Assessment in Undergraduate Medical Education». Articles. *Innovative Research Thoughts* 9 (4): 216–23. https://doi.org/10.36676/irt.2023-v9i4-029.

Schaper, Niclas. 2021. «Prüfen in der Hochschullehre». In *Handbuch Hochschuldidaktik*, herausgegeben von Robert Kordts-Freudinger, Niclas Schaper, Antonia Scholkmann, und Birgit Szczyrba, 87–102. Stuttgart: wbv Publikation. https://www.utb.de/doi/full/10.36198/9783838554082-87-102.

Sinharay, Sandip, Randy E. Bennett, Michael Kane, und Jesse R. Sparks. 2025. «Validation for Personalized Assessments: A Threats-to-Validity Approach». *Journal of Educational Measurement* n/a (n/a). https://doi.org/10.1111/jedm.12434.

Sireci, Stephen G. 2020. «Standardization and UNDERSTANDardization in Educational Assessment». *Educational Measurement: Issues and Practice* 39 (3): 100–5. https://doi.org/10.1111/emip.12377.

St-Onge, Christina, Meredith Young, Kevin W. Eva, und Brian Hodges. 2017. «Validity: One Word with a Plurality of Meanings». *Advances in Health Sciences Education: Theory and Practice* (Netherlands) 22 (4): 853–67. https://doi.org/10.1007/s10459-016-9716-3.

Tierney, Robin D. 2014. «Fairness as a multifaceted quality in classroom assessment». *Studies in Educational Evaluation* 43 (Dezember): 55–69. https://doi.org/10.1016/j.stueduc.2013.12.003.

Walker, Michael E., Margarita Olivera-Aguilar, Blair Lehman, Cara Laitusis, Danielle Guz-man-Orth, und Melissa Gholson. 2023. «Culturally Responsive Assessment: Provisio-nal Principles». *ETS Research Report Series* 2023 (1): 1–24. https://doi.org/10.1002/ets2.12374.

Wilcoxon, Frank. 1945. «Individual Comparisons by Ranking Methods». *Biometrics Bulletin* 1 (6): 80–8. JSTOR. https://doi.org/10.2307/3001968.

Wollny, Sebastian, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, und Hendrik Drachsler. 2021. «Are We There Yet? – A Systematic Literature Review on Chatbots in Education». *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.654924.

Anhang 1: Standardisierte Prüfung (SP)

Die TechSolutions GmbH steht vor der Herausforderung, eine umfassende KI-Strategie zu entwickeln, um ihre Wettbewerbsfähigkeit zu sichern. Wettbewerbsfähigkeit wird durch eine verbesserte Bestandsverwaltung im Lager, eine optimierte Routenplanung und einen personalisierten Kundenservice, z. B. alternative Lieferzeiten oder Vorschläge für ergänzende Produkte, hergestellt.

Für die Umsetzung steht ein Budget von 250.000 € zur Verfügung. Geplant ist eine schrittweise Implementierung, um Kosten zu kontrollieren. Gleichzeitig bestehen interne Bedenken: Mitarbeitende äussern die Sorge, dass Arbeitsplätze durch Automatisierung gefährdet sein könnten. Schulungen und Workshops wurden bisher nicht durchgeführt. Im Unternehmen gibt es bisher keine spezialisierten KI-Fachkräfte. Ein IT-Team mit grundlegenden Kenntnissen in Datenanalyse ist vorhanden.

Ein kritischer Punkt betrifft den Datenschutz: Sensible Daten wie Lieferadressen und Transportwege müssen im Einklang mit der DSGVO verarbeitet werden. Aktuell nutzt TechSolutions ein ERP-System sowie mehrere Datenbanken mit Lager- und Transportinformationen, die grundsätzlich für den Datenaustausch mit externen Lösungen geeignet sind.

Aufgabe:

Analysiere die Situation und entwickle eine Strategie, indem du die folgenden Teilaufgaben beantwortest.

1. Ist-Analyse und Zieldefinition:
- Beschreibe die aktuelle Situation der TechSolutions GmbH.
- Definiere mindestens drei Ziele für die Integration von KI.

2. Schlüsselressourcen und Technologien:
- Wähle KI-Technologien aus, um die Ziele zu erreichen und begründe deine Wahl.
- Benenne Ressourcen, die benötigt werden (z. B. Daten, Personal, Infrastruktur).

3. Konkrete Anwendungsszenarien:
- Entwickle zwei konkrete Anwendungsfälle, wie KI im Lager- und Transportmanagement eingesetzt werden kann.
- Benenne zwei Vorteile und zwei mögliche Herausforderungen dieser Anwendungsfälle.

4. Vorschlag zur Akzeptanzerhöhung:
- Skizziere einen Vorschlag, um die Akzeptanz für KI-Projekte bei den Mitarbeitenden und Stakeholdern zu erhöhen.

5. Chancen und Risiken:
- Analysiere die Chancen und Risiken der KI-Integration für TechSolutions GmbH.
- Erkläre 3 konkrete Massnahmen, um die Risiken zu minimieren.

Anhang 2: Kontextualisierte Prüfung (PG A)

Die MediCommerce AG, ein europaweit tätiger Online-Händler für digitale Medienprodukte, plant die Einführung einer KI-gestützten Plattform zur personalisierten Kundenberatung. Ziel ist es, mit Hilfe von Prompt Engineering und Data-Science-Analysen Kaufempfehlungen in Echtzeit zu geben, Retouren zu reduzieren und die Kundenzufriedenheit zu steigern.

Die technische Umsetzung soll in enger Zusammenarbeit mit einem IT-Team erfolgen, das bereits Grundkenntnisse in Datenanalyse besitzt. Ein Budget von 250.000 € ist vorgesehen. Die Unternehmensführung betont die Chancen für das Geschäftsmodell, gleichzeitig bestehen jedoch Bedenken bei Datenschutz, digitaler Selbstbestimmung der Kund:innen und möglichen rechtlichen Konflikten im Internet- und Medienrecht.

Kritische Punkte:

- Verarbeitung sensibler Kundendaten (Kaufhistorien, Zahlungsinformationen, Nutzungsprofile) unter DSGVO-Bedingungen.
- Risiko von algorithmischen Verzerrungen (Bias), die bestimmte Kundengruppen benachteiligen könnten.
- Keine etablierten Governance-Strukturen für KI-Management vorhanden.
- Erste Mitarbeitende äussern Skepsis bezüglich des «gläsernen Kunden» und ethischer Implikationen.

Aufgabe:

Analysiere die Situation und entwickle eine Strategie, indem du die folgenden Teilaufgaben beantwortest.

1. Ist-Analyse und Zieldefinition:

- Beschreibe die aktuelle Situation der MediCommerce AG.
- Definiere mindestens drei Ziele für eine ethisch verantwortbare Integration von KI.

2. Schlüsselressourcen und Technologien:

- Wähle KI-Technologien aus, um die Ziele zu erreichen und begründe deine Wahl.
- Benenne Ressourcen, die benötigt werden (z. B. Daten, Personal, Infrastruktur).

3. Konkrete Anwendungsszenarien:

- Entwickle zwei konkrete Anwendungsfälle, wie KI in der Kundenberatung und im Retourenmanagement eingesetzt werden kann.
- Benenne zwei Vorteile und zwei mögliche ethische oder rechtliche Herausforderungen dieser Anwendungsfälle.

4. Vorschlag zur Akzeptanzerhöhung:
- Skizziere einen Vorschlag, um die Akzeptanz für KI-Projekte bei den Mitarbeitenden und Stakeholdern zu erhöhen, unter besonderer Berücksichtigung ethischer Bedenken.

5. Chancen und Risiken:
- Analysiere die Chancen und Risiken der KI-Integration für MediCommerce AG.
- Erkläre 3 konkrete Massnahmen, um die Risiken zu minimieren.

Anhang 3: Argumentative Evidenzkette für F1

Theoriegeleitet		**Empirische Studie: Auswertungsdesign für F1**						
Aussage	Begründung	Annahme *(Fragebogenitems)*	Beleg *(Studienergebnisse)*					
			Gruppe	**MW**	**SD**	**g**	**W**	**p**
1. Personalisierte Prüfungen haben negative Auswirkungen auf die Prüfungssituation.	*Scoring: questionable score comparability:* Standardisierte Prüfungsaufgaben sind valider als personalisierte Prüfungsaufgaben, weil Leistungen so verglichen werden können.	Item 1: Die Prüfungsaufgabe kann mithilfe des Bewertungsbogens bewertet werden.	SP	4,22	1,30	0,00	75,50	> 0,05
			PG A	4,22	,44			
		Item 2: Der Bewertungsbogen ermöglicht eine standardisierte Auswertung der Prüfungsaufgabe.	SP	3,89	1,05	-0,10	82,50	> 0,05
			PG A	4,00	1,00			
2. Personalisierte Prüfungen haben positive Auswirkungen auf die Prüfungssituation.	*Administration: low motivation:* Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende motivierter sind diese zu lösen.	Item 3: Die inhaltliche Gestaltung der Prüfungsaufgabe wirkt motivierend auf den Lernenden.	SP	3,67	,87	-0,57	74,00	> 0,05
			PG A	4,11	,60			
		Item 4: Die Prüfungsaufgabe ist für eine berufliche Weiterbildung geeignet.	SP	4,22	,67	0,15	82,50	> 0,05
			PG A	4,11	,78			
	Administration: task or response not communicated: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende die zu erledigende Aufgabe besser verstehen.	Item 5: Die Aufgabenstellung der Prüfungsaufgabe ist verständlich.	SP	4,22	,67	-0,14	80,50	> 0,05
			PG A	4,33	,87			
		Item 6: Die Prüfungsaufgabe ist inhaltlich fehlerfrei.	SP	4,33	,50	0,11	79,50	> 0,05
			PG A	4,22	1,30			

	Evaluation: poor grasp of assessment information and its limitations: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende das Ergebnis leichter interpretieren können.	Item 7: Die Prüfungsaufgabe ermöglicht einen Rückschluss auf den aktuellen Lernstand des Lernenden.	SP	3,89	,60	0,00	83,00	> 0,05
			PG A	3,89	,93			
	Evaluation: inadequately supported construct interpretation: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil die Ergebnisse von personalisierten Prüfungen direkt der zu prüfenden Kompetenz zugeordnet werden können. Die Aufgabe verleitet nicht, individuelle Leistungen vorschnell auf umfassendere oder nicht intendierte Kompetenzbereiche zu verallgemeinern.	Item 8: Die Prüfungsaufgabe bezieht sich eindeutig auf die zu prüfende Kompetenz.	SP	4,33	,87	0,40	75,00	> 0,05
			PG A	4,00	,71			
		Item 9: Die Prüfungsaufgabe bezieht sich ausschliesslich auf die zu prüfende Kompetenz.	SP	3,56	,73	0,84	66,50	> 0,05
			PG A	2,89	,78			
	Evaluation: biased interpretation or explanation: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie die Interpretation der Ergebnisse durch ihre erhöhte Anschlussfähigkeit an den individuellen Lernweg der geprüften Person nachvollziehbarer machen.	Item 10: Die Prüfungsaufgabe schafft Anknüpfungspunkte für die individuelle Weiterentwicklung des Lernenden.	SP	4,00	,87	0,27	76,00	> 0,05
			PG A	3,78	,67			
	Decision: poor pedagogical decisions: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie stärker auf den einzelnen Lernenden abgestimmt sind und das Geben von stärkeorientierten Feedback ermöglichen.	Item 11: Die Prüfungsaufgabe passt zum beruflichen Hintergrund des Lernenden.	SP	2,89	,93	-1,36	58,00	< 0,05
			PG A	4,11	,78			
		Item 12: Die Prüfungsaufgabe ermöglicht eine Rückmeldung, die die Leistung des Lernenden anerkennend würdigt.	SP	4,00	,87	0,22	82,50	> 0,05
			PG A	3,78	1,09			

	Impact: positive consequences not achieved: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie gezieltere Rückmeldungen ermöglichen, Motivation fördern und relevanter für die zukünftige Anwendung von Kompetenzen wahrgenommen werden.	Item 13: Die Prüfungsaufgabe erlaubt eine Rückmeldung, die gezielt auf die Stärken und Schwächen des Lernenden eingeht.	SP	3,89	,78	0,26	81,00	> 0,05
			PG A	3,67	,87			
		Item 14: Die Prüfungsaufgabe ermöglicht eine passgenaue Empfehlung für den weiteren Bildungsweg des Lernenden.	SP	2,67	,87	0,00	85,50	> 0,05
			PG A	2,67	,87			
	Impact: serious negative impact occurs: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie als fairer wahrgenommen werden.	Item 15: Aus Sicht eines Prüfenden ist die Prüfungsaufgabe fair gestaltet.	SP	4,00	,87	0,00	84,00	> 0,05
			PG A	4,00	1,00			
		Item 16: Die Prüfungsaufgabe ist geeignet, das Lern- und Kompetenzziel des Kurses zu überprüfen.	SP	4,22	,83	0,15	81,00	> 0,05
			PG A	4,11	,60			
		Item 17: Die Prüfungsaufgabe bezieht die Lerninhalte des Kurses ein.	SP	4,22	,83	0,23	81,00	> 0,05
			PG A	4,00	1,00			
		Item 18: Die Prüfungsaufgabe erfordert vom Lernenden das Anwenden gelernter Inhalte.	SP	4,56	,53	0,56	74,50	> 0,05
			PG A	4,11	,93			
		Item 19: Die Prüfungsaufgabe beinhaltet Informationen, die ausschliesslich prüfungsrelevant sind.	SP	3,89	,60	0,64	71,50	> 0,05
			PG A	3,33	1,00			
		Item 20: Die Prüfungsaufgabe kann für alle Lernenden des Kurses eingesetzt werden, ohne dass dadurch Vor- oder Nachteile für bestimmte Lernendengruppen entstehen.	SP	3,33	1,00	0,19	82,00	> 0,05
			PG A	3,11	1,27			

Anhang 4: Vorgelagerte Abfrage für F2

Ziel	Frage	Auswahlmöglichkeiten	Weiterleitung
Vorab-Abfrage der generellen Nutzung des Chatbots	1. Haben Sie den Chatbot genutzt?	• Ja. • Nein.	• Wenn «Ja» ausgewählt wird, erfolgt eine automatische Weiterleitung zu Frage 2. • Wenn «Nein» ausgewählt wird, erfolgt eine automatische Weiterleitung zu Frage 3.
Vorab-Abfrage der Intensivität der Nutzung des Chatbots	2. Wie viele Anfragen haben Sie an den Chatbot gestellt?	• weniger als 5 • zwischen 5–10 • zwischen 11–15 • zwischen 16–20 • mehr als 20	• Unabhängig von der Auswahl wird bei der Auswahl einer Möglichkeit zur ersten inhaltlichen Frage zum Chatbot (siehe Anhang 5) weitergeleitet.
Vorab-Abfrage der Gründe für die Nicht-Benutzung des Chatbots	3. Begründen Sie, weshalb Sie den Chatbot nicht ausprobiert haben.	*Freitextfeld*	• Nach Eingabe eines oder mehrerer Gründe erfolgt die Beendigung des Fragebogens.

Anhang 5: Argumentative Evidenzkette für F2

Theoriegeleitet		**Empirische Studie: Auswertungsdesign für F2**							
Aussage	Begründung	Annahme *(Fragebogenitems)*	Beleg *(Studienergebnisse)*						
			Berechnungen		5er Likert-Skala				
		Benutzt der Lernende während der Prüfungssituation den Chatbot, …	MW	SD	1	2	3	4	5
1. Personalisierte Prüfungen haben negative Auswirkungen auf die Prüfungssituation.	*Scoring: questionable score comparability:* Standardisierte Prüfungsaufgaben sind valider als personalisierte Prüfungsaufgaben, weil Leistungen so verglichen werden können.	Item 1: … kann die standardisierte Auswertung der Prüfung gefährdet werden.	2,35	1,22	4	8	1	3	1
		Item 2: … kann die Vergleichbarkeit der Prüfungsleistung des Lernenden zur Leistung eines anderen Lernenden eingeschränkt werden.	3,06	1,35	2	6		7	2
		Item 3: … kann die Bewertung nicht ausschliesslich die Eigenleistung des Lernenden widerspiegeln.	3,35	1,37	1	5	3	3	5

2. Personalisierte Prüfungen haben positive Auswirkungen auf die Prüfungssituation.	*Administration: low motivation:* Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende motivierter sind diese zu lösen.	Item 4: … kann der Lernende bei der Bearbeitung der Prüfungsaufgabe motiviert werden.	4,00	,94		1	4	6	6
	Administration: task or response not communicated: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende die zu erledigende Aufgabe besser verstehen.	Item 5: In einer beruflichen Weiterbildung ist die Benutzung dieses Chatbots während der Prüfungssituation legitim.	3,47	,87		3	4	9	1
		Item 6: … kann der Lernende beim Verstehen der Prüfungsaufgabe unterstützt werden.	4,18	1,02	1	1		8	7
		Item 7: … kann der Lernende fehlerhafte Informationen vom Chatbot erhalten.	2,65	1,22	2	8	3	2	2
	Evaluation: poor grasp of assessment information and its limitations: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil Lernende das Ergebnis leichter interpretieren können.	Item 8: … kann der tatsächliche Lernstand des Lernenden sichtbar gemacht werden.	2,82	,95	1	6	5	5	
	Evaluation: inadequately supported construct interpretation: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil die Ergebnisse von personalisierten Prüfungen direkt der zu prüfenden Kompetenz zugeordnet werden können. Die Aufgabe verleitet nicht, individuelle Leistungen vorschnell auf umfassendere oder nicht intendierte Kompetenzbereiche zu verallgemeinern.	Item 9: … kann das Prüfungsziel für den Lernenden erkennbar gemacht werden.	3,53	1,07	1	2	3	9	2

	Evaluation: biased interpretation or explanation: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie die Interpretation der Ergebnisse durch ihre erhöhte Anschlussfähigkeit an den individuellen Lernweg der geprüften Person nachvollziehbarer machen.	Item 10: … kann die Planung des weiteren Lernwegs unterstützt werden.	4,12	,86		1	2	8	6
	Decision: poor pedagogical decisions: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie stärker auf den einzelnen Lernenden abgestimmt sind und das Geben von stärkeorientierten Feedback ermöglichen.	Item 11: … können die individuellen Bedürfnisse des Lernenden während der Prüfung adressiert werden.	3,88	,86		1	4	8	4
		Item 12: … kann der Lernende nützliches Feedback während der Prüfung erhalten.	4,06	,90		1	3	7	6
	Impact: positive consequences not achieved: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie gezieltere Rückmeldungen ermöglichen, Motivation fördern und relevanter für die zukünftige Anwendung von Kompetenzen wahrgenommen werden.	Item 13: … kann der Lernende nützliches Feedback nach der Prüfung erhalten.	3,35	,86		2	9	4	2
		Item 14: … kann eine passgenaue Empfehlung für den weiteren Bildungsweg des Lernenden gegeben werden.	3,47	1,01		4	3	8	2

	Impact: serious negative impact occurs: Personalisierte Prüfungsaufgaben sind valider als standardisierte Prüfungsaufgaben, weil sie als fairer wahrgenommen werden.	Item 15: Aus Sicht eines Prüfenden ist die Benutzung dieses Chatbots in einer Prüfungssituation fair.	3,59	1,06		3	5	5	4
		Item 16: … kann die Überprüfung, ob der Lernende das Lern- und Kompetenzziels tatsächlich erreicht, gefährdet werden.	2,76	1,09	1	7	6	1	2
		Item 17: … kann das Anforderungsniveau der Prüfungsaufgabe verändert werden.	3,88	1,05		2	4	5	6
		Item 18: … kann der Lernende durch nicht prüfungsrelevante Informationen abgelenkt werden.	2,59	1,23	2	9	2	2	2
		Item 19: … kann der Lernende einen inhaltlichen Vorteil bei der Bearbeitung der Prüfungsaufgabe erhalten.	3,12	1,50	4	1	5	3	4
		Item 20: … können Vorteile für technikaffine Lernende entstehen.	3,47	1,33	1	4	3	4	5
		Item 21: … ist die Kommunikation neutral und diskriminierungsfrei.	4,06	,66		1		13	3