Lernkultur durch innovative Prüfungsformate entwickeln

Kerstin Kusch, Tobias Weber, Claudia Albrecht

Abstract: *Die Prüfungstheke (Albrecht et al., 2023) ermöglicht Studierenden ihre Prüfung und damit das Lernen selbst zu gestalten. Das Konzept wurde mit Befragungen in einer Vorlesung mit 142 Studierenden evaluiert. Dreimal konnten die Studierenden zwischen Onlinetestat und vertiefenden Portfolioaufgaben wählen. Die Evaluation zeigt, dass die Prüfungstheke in großen Studierendengruppen umsetzbar ist, akzeptiert wird und Studierende die Freiräume wahrnehmen, nutzen und schätzen. Verglichen mit Testaten gehen Portfolioaufgaben bei gleichen Prüfungsergebnissen mit verteiltem und vertieftem Lernen einher. So trägt die Prüfungstheke zur Entwicklung der Lern- und Prüfungskultur bei mit Weiterentwicklungspotenzial durch lernförderliches Feedback, Reflexionsunterstützung und KI-gestützten Lerngelegenheiten.*

The examination kiosk (Albrecht et al., 2023) enables students to organize their examinations and thus their learning themselves. The concept was evaluated through surveys in a lecture with 142 students, who could choose between online tests and in-depth portfolio tasks on three occasions. The evaluation shows that the examination kiosk can be implemented in large groups, is accepted, and students appreciate the freedom to choose. Compared to tests, portfolio tasks are associated with distributed and in-depth learning while the examination results are equal. Therefore, the examination kiosk contributes to the development of the learning and examination culture. It will be further developed through feedback that promotes learning, reflection support, and AI-supported learning opportunities.

Keywords: *Prüfungstheke; Portfolioprüfungen; Selbstgesteuertes Lernen; Prüfungskultur; Lernaktivitäten; generative Künstliche Intelligenz; examination kiosk; portfolio examinations; self-directed learning; examination culture; learning activities; generative artificial intelligence*

1. Einleitung

Im Rahmen des Projektes »Portfolioprofis: Portfolioprüfungen als Initialzündung einer veränderten Lern- und Lehrkultur« (gefördert durch die Stiftung Innovation in der Hochschullehre im Rahmen der Förderlinie »Freiraum 2022«) werden u.a. neue Prüfungsformate entwickelt, eingesetzt und evaluiert, die es auch in großen Studierendengruppen ermöglichen sollen, kompetenzorientiert und studierendenzentriert zu prüfen und selbstgesteuerte Lernprozesse zu unterstützen. Entstanden ist dabei die Prüfungstheke, die den Studierenden die Möglichkeit bietet, »innerhalb eines vorgegebenen inhaltlichen und methodischen Rahmens den Inhalt und das Format ihrer Prüfungsleistung selbstständig zu wählen. Als benotete Portfolio-Prüfung besteht sie aus mehreren Teilleistungen« (Albrecht et al., 2023).

Das didaktische Konzept der Prüfungstheke betont die selbstgesteuerte Gestaltung der Prüfung und damit den Lernprozess der Studierenden. Dabei werden Studierenden Entscheidungs- und Handlungsspielräume eröffnet, die auf den neun Dimensionen des selbstgesteuerten Lernens nach Dyrna (2021) basieren. Beim Einsatz der Prüfungstheke können daher Studierende grundsätzlich über Lernziele, Lerninhalte, Lernquellen, Lernmethoden, Lernweg, Lerneinschätzung, Lernpartner, Lernzeit und Lernort selbst bestimmen (Albrecht et al., 2023).

Das Rahmenkonzept der Prüfungstheke wurde auf die Anforderungen der Vorlesung Ingenieurpsychologie des Bachelorstudienganges Psychologie an der Technischen Universität Dresden angepasst und dort im Sommersemester 2023 erstmalig eingesetzt. An der Lehrveranstaltung nahmen 142 Studierende teil. Die Studierenden konnten zu drei Zeitpunkten im Semester entscheiden, ob sie die in dem entsprechenden Zeitraum behandelten Inhalte in Form eines online-Testates oder durch die vertiefende Bearbeitung selbstgewählter Inhalte durch die Abgabe zweier Portfolioaufgaben prüfen lassen wollen. Dabei war vorgeschrieben, dass mindestens einmal Portfolioaufgaben eingereicht werden sollen. Die Portfolioaufgaben waren eng an im Lernprozess entstehende Produkte angelehnt. Studierende konnten entweder eine fachspezifische Beispielstudie planen, eine Concept-Map erstellen, ein Exzerpt verfassen, Prüfungsfragen formulieren oder auf die Vorlesung bezogene Diskussionsfragen beantworten (Albrecht et al., 2023).

Der Einsatz der Prüfungstheke wurde begleitend evaluiert. Um die Wahrnehmung der Prüfungstheke und Lernaktivitäten zu untersuchen, wurden die Studierenden nach jedem der drei Abgabezeitpunkte und nach dem Testat mithilfe eines Online-Fragebogens befragt. Nach Abschluss des Semesters wurden die Studierenden nochmals befragt, um herauszufinden, wie sie die Möglichkeiten zur selbstgesteuerten Gestaltung ihres Lernprozesses nutzten. Mit diesem Beitrag werden die Ergebnisse der empirischen Untersuchung zum Einsatz der Prüfungstheke zusammengefasst und diskutiert, inwiefern die Prüfungstheke kompetenzorientiertes Prüfen ermöglicht und zur Entwicklung der Lern- und Prüfungskultur beiträgt.

Abschließend werden die Auswirkungen des Einbezuges von Künstlicher Intelligenz (KI) als Unterstützung der Aufgabenbearbeitung diskutiert.

2. Theorie

Prüfungen sind integraler Bestandteil der Hochschullehre und erfüllen verschiedene Funktionen (Schaper, 2021). Auf Seiten der Studierenden hat die Prüfungsgestaltung einen Einfluss auf das Lernverhalten (Baird et al., 2017). Darüber hinaus können Prüfungen selbst als Lerngelegenheiten angesehen werden. Dies rückt jedoch durch die starke Betonung von summativen Prüfungen mit Rechtsfolgen (meist in Form einer Modulprüfung) und die damit einhergehenden Konsequenzen (Weiterkommen im Studium vs. Wiederholung oder gar Exmatrikulation) zumeist in den Hintergrund, was die lernförderlichen Potenziale von Prüfungen minimiert (Reinmann, 2022). Die beschriebene Situation ist das Ergebnis des Zusammenwirkens von Faktoren auf verschiedenen Ebenen, darunter der rechtlichen-organisatorischen Ebene (z.B. Studiendokumente), der technischen Ebene (z.B. die Umsetzbarkeit bestimmter digitaler Durchführungsszenarien) sowie der didaktischen Ebene (z.B. die Konzeption mehr oder weniger kompetenzorientierter Prüfungsformate). Wie sähe eine neue Prüfungskultur aus und welche Bedingungen müssen sich auf diesen Ebenen ändern, um diesen Zustand zu erreichen? Das Hochschulforum Digitalisierung skizzierte vor Kurzem eine »Vision einer neuen Prüfungskultur« (Budde et al., 2024). Auf der Ebene der Prüfungsszenarien und Prüfungsformate arbeiten die Autor:innen folgende Punkte als Bestandteile der neuen Prüfungskultur heraus: Prüfungen werden als Teil des Lernprozesses verstanden. Es gibt für jeden Studiengang lernzielorientierte Kriterienraster. Der Fokus von Prüfungen liegt auf Kompetenzen und nicht auf reinen Wissensabfragen, was einen starken Anwendungsbezug sowie eine Handlungs- und Problemorientierung erfordert. Transformative, technologische und digitale Schlüsselkompetenzen sind sinnvoll in Prüfungen integriert.

Zur Einordnung unseres Vorhabens sind aber vor allem zwei Aspekte wichtig, die ebenfalls als Kernelemente einer neuen Prüfungskultur angesehen werden: Studierende können selbst darüber entscheiden, welches Prüfungsformat sie wählen und welchen inhaltlichen Fokus sie setzen. Dies sollte mit dem Ziel der eigenen Kompetenzentwicklung geschehen. Heterogenität und Diversität der Studierenden werden durch eine hohe Flexibilität Rechnung getragen, gleichzeitig werden Fairness und Chancengleichheit berücksichtigt.

Diese Aspekte sind der innovative Kern der Prüfungstheke. Die Wahlmöglichkeit sowie die Differenziertheit der Prüfungsleistungen (Concept-Map, Testat, Prüfungsfragen erstellen, Reflexionspaper) erlauben es Studierenden zeitlich flexibel zu agieren, für sich ein Anspruchsniveau der Prüfungsleistung zu definieren, eigene

Lernstrategien an das gewählte Prüfungsziel anzupassen, sowie individuell erachtete »attraktive« Prüfungsformate auszusuchen, die je nach Themenkomplex den eigenen Lernprozess fördern können. Damit wurden mit der Prüfungstheke die verschiedenen Ebenen (kognitive Ebene, metakognitive Ebene, Ebene des Selbst) des selbstregulierten Lernens adressiert, die in ihrer Gesamtheit das System »Lernender« definieren (Boekaerts, 1999).

Da die Prüfungstheke erstmals praktisch erprobt wurde und Wahlfreiheit einerseits durch Studierende erwünscht ist (Perrella et al., 2015), andererseits mit Kosten einhergeht (Iyengar & Lepper, 2000; Smith & Walker, 1993), stellte sich die Frage: Wie wird die Prüfungstheke durch die Studierenden akzeptiert?

Die Prüfungstheke als Portfolioprüfung fungiert auch als Assessment mit Rechtsfolgen, so dass sie objektiv, zuverlässig (reliabel), gültig (valide), ökonomisch und fair sein muss (vgl. Schaper, 2021). Daher wurde überprüft, wie sich die Güte zwischen Portfolioaufgaben und traditionellen Testaten mit Multiple-Choice-Aufgaben unterscheidet.

Selbstgesteuertes Lernen wird durch die Prüfungstheke indirekt (Friedrich & Mandl, 1997) über die eingeräumten Gestaltungsspielräume gefördert. Darum wurde mit der Evaluation geprüft, wie die Studierenden die sich ihnen eröffnenden Freiräume wahrnehmen, nutzen und welche Elemente sie als förderlich oder hinderlich für den Lernprozess empfinden.

3. Methode

3.1 Stichprobe

Im Sommersemester 2023 waren 142 Personen in die Vorlesung Ingenieurpsychologie an der TU Dresden eingeschrieben. Davon studierten 125 Bachelor Psychologie und 17 einen anderen Studiengang. Alle eingeschriebenen Studierenden wurden per E-Mail zur Befragung eingeladen. Insgesamt nahmen 73 Personen an den Evaluationsbefragungen teil. Davon waren 68 jünger als 25 Jahre und 5 Personen älter; 68 Teilnehmende studieren den Bachelor Psychologie und 5 Personen einen anderen Studiengang. Die Studierenden wollten mit dem Besuch der Lehrveranstaltung Wissen erwerben (n = 31, bspw. »Wissen erweitern«), das Modul abschließen (n = 18, wie »Sehr gute Noten« oder »Durchkommen, gut bestehen, abhaken«) oder verfolgten kein (n = 13) oder andere Lernziele (n = 8, wie »Interesse wecken«). Im Durchschnitt waren sie wenig an den Inhalten der Ingenieurpsychologie interessiert.

3.2 Ablauf

Die Daten wurden online mit LimeSurvey 3.27 im Zeitraum vom 17.05. bis zum 31.10.2023 erhoben. Es gab fünf Erhebungszeitpunkte: jeweils zu den drei Abgabezeitpunkten, nach dem Testat und nachdem die Ergebnisse rückgemeldet wurden. Die Studierenden füllten die ersten 4 Befragungen innerhalb von 1 bis 6 Tagen aus. Dafür benötigten sie durchschnittlich zwischen 3 und 11 Minuten. Die abschließende Befragung wurden innerhalb von 33 Tagen durchgeführt und dauerte durchschnittlich 21 Minuten. Dafür erhielten die Teilnehmenden 10 Euro oder eine Versuchspersonenstunde als Aufwandsentschädigung.

Die erste Befragung füllten 45 von 84 Personen, die Portfolioaufgaben abgaben, vollständig aus (54 %), 6 von 35, die ein Testat schrieben (17 %) und 7 von 23, die keine Prüfungsleistung erbrachten (30 %). An der zweiten Befragung nahmen 26 von 96 Personen mit Aufgaben (27 %), 2 von 21 mit Testat (10 %) und 1 von 35 ohne Prüfungsleistung (4 %) vollständig teil. Nach der dritten Entscheidung beantworteten 17 von 91 Personen mit Aufgaben (19 %), 2 von 26 mit Testat (8 %) und keine von 25 ohne Prüfungsleistung vollständig die Umfrage. Von den 53 Personen, die ein Testat mitschrieben, nahmen 8 (15 %) an der Befragung nach dem Testat teil. An der Abschlussbefragung beteiligten sich 61 der 117 Personen, die alle Leistungen der Prüfungstheke absolvierten (52 %).

Zu jedem Befragungszeitpunkt wurden die gewählte Prüfungsleistung und bei Bedarf Angaben zur Person erfasst. Während des Semesters wurden in den Befragungen Aufwand, Auswahlkriterien, investierte Zeit und Wohlbefinden erhoben. Gründe, keine Prüfungsleistung abzulegen, wurden vom ersten bis zum dritten Zeitpunkt erfasst. Lernaktivitäten wurden nach der zweiten Entscheidung und nach dem Testat erfragt. Zufriedenheit wurde zu Beginn und am Ende der Evaluation bestimmt.

Die wahrgenommenen und genutzten Gestaltungsspielräume nach Dyrna (2021) und die Sicht auf die weitere Nutzung der Prüfungstheke wurden in der abschließenden Erhebung erfragt.

3.3 Instrumente

Zur Auswertung lagen die Prüfungsergebnisse (Punktwerte pro gewählter Prüfungsleistung) vor. Diese waren mit dem Studiengang der Person verknüpft. Für die Befragungen wurden standardisierte oder ad hoch erzeugte Fragen verwendet. Fragen, die die Auswahl der Prüfungsleistung betreffen, werden hier nicht berichtet.

Zufriedenheit wurde mit der Frage »Im Großen und Ganzen, wie zufrieden sind Sie mit der Prüfungstheke, also der Möglichkeit, verschiedene Teilleistungen im Rahmen der Portfolioprüfung auszuwählen?« auf einer fünfstufigen Likert-Skala

erfasst (von 1 = sehr zufrieden bis 5 = überhaupt nicht zufrieden). In der ersten Befragung erhielten sehr und überwiegend zufriedene Studierende die Frage: »Was gefällt Ihnen an der Prüfungstheke besonders gut und sollte beibehalten werden?«, waren die Studierenden weniger zufrieden, erhielten sie die Frage: »Was sollten wir an der Prüfungstheke Ihrer Meinung nach ändern?«.

Angaben zur Person: Die Studierenden gaben an, ob sie Bachelor Psychologie studierten, jünger als 25 Jahre waren oder älter, ihr persönliches Lernziel (Item aus Frank, o.J.), sowie ihr Fachinteresse mit der Aussage »Bereits vor Besuch der Lehrveranstaltung interessierten mich die Inhalte der Ingenieurpsychologie.« auf einer fünfstufigen Likert-Skala (1 = gar nicht zutreffend bis 5 = völlig zutreffend).

Gründe, keine Prüfungsleistung zu erbringen: Studierende, die keine Leistung erbrachten und sich nicht für ein Testat anmeldeten, wurden gefragt: »Wie kam es dazu, dass Sie keine Teilleistung auswählten?«. Dazu konnte aus sechs Antwortoptionen, wie »Mich hindern persönliche Gründe.«, und der Option »Sonstiges« mit freier Antwort gewählt werden.

Lernaktivitäten wurden mit der Kurzskala des Fragebogens Lernstrategien im Studium (LIST-K; Klingsieck, 2018) erhoben. In dieser Evaluation wurden die Fragen auf den Zeitraum, in dem die Aufgaben bearbeitet oder für das Testat gelernt wurde, beschränkt. Die Lernstrategien werden auf 13 Subskalen mit jeweils 3 Items erfasst. Die Häufigkeit verschiedener Tätigkeiten, wie bspw. »Ich denke mir konkrete Beispiele zu bestimmten Lerninhalten aus.«, wird auf einer fünfstufigen Antwortskala von 1 = *sehr selten* bis 5 = *sehr oft* eingeschätzt. Erfasst werden die kognitiven Strategien »Organisieren« ($\alpha = .20$), »Elaborieren« ($\alpha = .70$), »Kritisches Prüfen« ($\alpha = .74$) und »Wiederholen« ($\alpha = .81$), die metakognitiven Strategien »Planen« ($\alpha = .63$), »Kontrollieren« ($\alpha = .79$) und »Regulieren« ($\alpha = .72$), die Strategien zum Management interner Ressourcen »Aufmerksamkeit« ($\alpha = .91$), »Anstrengung« ($\alpha = .34$) und »Zeitmanagement« ($\alpha = .77$) sowie die Strategien zum Management externer Ressourcen »Lernen mit Studienkolleg_innen« ($\alpha = .83$), »Literatur« ($\alpha = .80$) und »Lernumgebung« ($\alpha = .46$).

Die wahrgenommenen und genutzten Entscheidungs- und Handlungsspielräume wurden über acht der neun Dimensionen selbstgesteuerten Lernens (Dyrna, 2021; eine Dimension wurde versehentlich weggelassen) erfasst. Die erhobenen Dimensionen waren »Lernziele festlegen«, »Lerninhalte auswählen«, »Lernmethoden festlegen«, »Lernmaterialien und -medien auswählen«, »Lernzeiten festlegen«, »Lernort bzw. Lernumgebung festlegen«, »Lernergebnisse bewerten und/oder reflektieren« und »Lernpartner:innen auswählen«. Die Wahrnehmung der Spielräume wurde über die Frage: »Wie sehr ermöglichte Ihnen die Prüfungstheke grundsätzlich, den Lernprozess hinsichtlich der folgenden Aspekte selbst zu gestalten? Antworten Sie bitte unabhängig davon, ob Sie die Möglichkeit gut fanden oder ob Sie sie nutzten.« Die Nutzung der Spielräume wurde mit der Frage »In welchem Ausmaß gestalteten Sie folgende Aspekte des Lernprozesses selbst?«

erhoben. Antworten konnten die Studierenden auf einer fünfstufigen Skala (von 1 = nicht bis 5 = sehr) oder der Antwortoption nicht zutreffend/weiß nicht. Herausforderungen wurden mit der Frage »Wie herausfordernd fanden Sie im Rahmen Ihres Lernprozesses die Gestaltung der folgenden Aspekte?« und hilfreiche Aspekte mit der Frage »Rückblickend betrachtet – wie hilfreich für Ihren eigenen Lernprozess fanden Sie die folgenden Aspekte?« ermittelt.

Wie die Prüfungstheke den Lernprozess beeinflusste und welche Unterstützungsbedarfe die Studierenden wahrnahmen, wurde mit 13 Aussagen (z.B. »Ich habe einzelne Themen vertieft bearbeitet.«) auf einer fünfstufigen Skala (von 1 = stimme überhaupt nicht zu bis 5 = stimme voll und ganz zu) oder der Option nicht zutreffend/weiß nicht erfasst. Auf der gleichen Skala wurde mit fünf Aussagen, wie »Die Prüfungstheke sollte in der Vorlesung Ingenieurpsychologie wieder eingesetzt werden« bestimmt, welche zukünftigen Einsatzmöglichkeiten die Studierenden sehen.

3.4 Auswertung

Die statistischen Analysen wurden mit R, Version 4.4.0 (R Core Team, 2024) durchgeführt. Für die schließende Statistik wurden optimale Stichprobenumfänge in A-priori-Teststärkenanalysen mit pwr, Version 1.3-0 (Champely, 2020) bzw. TOSTER, Version 0.8.2 (Caldwell, 2022; Lakens, 2017) berechnet. Festgelegt wurden ein Signifikanzniveau von 5 %, Teststärke von $1–\beta = .80$ und eine Effektstärke von $d = .4$ bzw. $r = .2$ für zweiseitige Tests. Diese Größe gilt als erwartbarer und praktisch bedeutsamer Wert für psychologische Fragestellungen (Brysbaert, 2019).

4. Ergebnisse

4.1 Akzeptanz der Prüfungstheke

Von den 142 in die Vorlesung eingeschriebenen Personen absolvierten 117 Studierende (82 %) ihre Prüfungsleistungen im Rahmen der Vorlesung. Von diesen wählten 88 Personen die Portfolioaufgaben öfter als das vorgeschriebene eine Mal. Mit einem Chi-Quadrat-Anpassungstest wurde geprüft, ob der Anteil an Personen, die Aufgaben häufiger wählten als vorgeschrieben, gleich 0.1 ist. Alternative Aufgaben wurden signifikant häufiger gewählt als durch die Regularien vorgegeben ($\chi^2(1, N = 117) = 552.87$, $p < .001$).

Sowohl zur ersten als auch zur letzten Befragung waren die Studierenden mit der Prüfungstheke überwiegend zufrieden ($M_{t1} = 1.97$, $SD_{t1} = 0.84$, $n_{t1} = 58$; $M_{t5} = 1.69$, $SD_{t5} = 0.56$, $n_{t5} = 61$). Von 32 Personen lagen die Zufriedenheitswerte für beide Zeitpunkte vor. Deren Zufriedenheit stieg etwas vom ersten ($M_{t1} = 1.91$, $SD_{t1} = 0.73$)

zum letzten Erhebungszeitpunkt (M_{t5} = 1.62, SD_{t5} = 0.55, wobei geringere Werte eine höhere Zufriedenheit ausdrücken). Auf eine inferenzstatistische Prüfung wurde verzichtet, da der optimale Stichprobenumfang von n = 52 nicht erreicht wurde. Personen, die sich in der ersten Befragung eher zufrieden zeigten, gefielen die Freiheit und Vielfalt der Auswahlmöglichkeiten (n = 22), die Möglichkeit, die Arbeitsbelastung im Semester zu regulieren (n = 7) und der Umstand, dass Testat und Aufgaben alternativ angeboten wurden (n = 7). Weniger zufriedene Studierende empfanden Aufwand und Anforderungen der Aufgaben als veränderungsbedürftig (n = 3).

Als Gründe, sich nicht bewerten zu lassen, wurden am häufigsten gewählt: die Arbeitsbelastung im Semester (n = 6), persönliche Gründe, wie Erkrankungen oder fehlende Kinderbetreuung (n = 6) und einmal das Umentscheiden für ein anderes Modul. Durch die Prüfungstheke bedingte Gründe wurden nicht benannt.

4.2 Eignung der Prüfungstheke als summatives Assessment

Die Spannweite der möglichen Punktwerte unterschied sich zwischen Portfolioaufgaben (von 1 bis 10) und Testaten (von 0 bis 10). Daher wurden die Punktwerte anhand der theoretischen Spannweiten normalisiert, wobei 0 Punkte den geringsten und 1 Punkt den höchsten zu erreichenden Wert widerspiegeln. Insgesamt erzielten die Studierenden Ergebnisse zwischen 0.45 und 1.00 Punkten (M = 0.89, SD = 0.08).

Die Retest-Reliabilität wurde als Intraklassenkorrelation mit dem Paket psych, Version 2.4.3 (Revelle, 2024) getrennt für Portfolioaufgaben und Testate berechnet. Jeweils wurden mit einem 2-Wege-Modell mit gemischten Effekten die einzelnen normalisierten Punkte auf absolute Übereinstimmung geprüft. Die Retest-Reliabilität der Portfolioaufgaben (F(87,174) = 2.2, p < .001, ICC_2 = 0.24, 95 % KI [0.19,0.29]) ist als mangelhaft und die der Testate (F(28,56) = 12, p < .001, ICC_2 = 0.78, 95 % KI [0.74,0.82]) als moderat bis gut einzustufen (Koo & Li, 2016).

Um eine mindestens angemessene konvergente Validität von r = .55 (Evers et al., 2013) zu überprüfen, war ein Stichprobenumfang von n = 23 optimal. Da sich 53 Studierende sowohl mit Portfolioaufgaben als auch mit einem Testat prüfen ließen, wurde die Analyse durchgeführt. Die Bewertungsergebnisse der Portfolioaufgaben und der Testate korrelierten mit r = .35, 95 % KI [.09, .57], t(51) = 2.66, p = .010, was für unzureichende bis angemessene konvergente Validität spricht (Evers et al., 2013).

Für die 53 Fälle, in denen sowohl Werte für Testate (M = 0.87, SD = 0.12) als auch Portfolioaufgaben (M = 0.90, SD = 0.06) vorlagen, wurde ein Äquivalenztest (TOST) mit dem R-Paket TOSTER, Version 0.8.2 (Caldwell, 2022; Lakens, 2017) berechnet. Als Wert für praktische Gleichheit der Ergebnisse wurde 0.1 angenommen (10-prozentige Abweichung). Der optimale Stichprobenumfang lag bei 36, so dass der Äquivalenztest durchgeführt wurde. Sowohl der Äquivalenztest als auch der Unterschiedstest (t(52) = −4.66, p < .001 (ΔM = 0.03 90 % KI [-0.00, 0.05]; Hedges'g(z)

= 0.225 90 % KI [-0.001, 0.45])) weisen darauf hin, dass sich die Punkte nicht signifikant voneinander unterscheiden.

Für die Bearbeitung der Portfolioaufgaben inklusive der besuchten Vorlesungszeit benötigten die Studierenden zum ersten und zweiten Zeitpunkt durchschnittlich 15 (SD_{t1} = 7.7, n_{t1} = 44, ein Wert von 120 wurde als Ausreißer entfernt; SD_{t2} = 6.2, n_{t2} = 27) und zum dritten Zeitpunkt 13 Stunden (SD = 4.5, n = 17). Durchschnittlich bereiteten sich die Studierenden auf ein Testat inklusive der besuchten Vorlesungszeit zum ersten Zeitpunkt 14 (SD = 6.2, n = 7), zum zweiten Zeitpunkt 20 (SD = 21.2, n = 2), zum dritten Zeitpunkt 29 Stunden (SD = 32.4, n = 3) und laut der Befragung nach dem Testat 17 Stunden pro Testat (SD = 5.1, n = 9) vor. Aufgrund der geringen Rücklaufquote v.a. der Studierenden, die ein Testat schrieben, wird kein inferenzstatistischer Test berechnet

Mit dem LIST-K konnten Lernstrategien für 23 Personen, die Aufgaben einreichten und 8 Personen, die ein Testat schrieben, bestimmt werden. Tendenziell gaben Personen, die eine Aufgabe einreichten, an, häufiger elaboriert und kritisch geprüft zu haben als Personen, die ein Testat schrieben (Elaborieren: $M_{\text{Aufgabe}.}$ = 4.0, SD_{Aufgabe} = 0.64 vs. M_{Testat} = 3.1, SD_{Testat} = 1.26; Kritisches Prüfen: $M_{\text{Aufgabe}.}$ = 3.6, SD_{Aufgabe} = 0.80 vs. M_{Testat} = 2.3, SD_{Testat} = 1.01) Dagegen berichteten Personen, die ein Testat schrieben, häufiger zu wiederholen (M_{Testat} = 2.4, SD_{Testat} = 1.11 vs. M_{Aufgabe} = 1.3, SD_{Aufgabe} = 0.44), mit Kommiliton:innen zu lernen (M_{Testat} = 3.8, SD_{Testat} = 0.89 vs. $M_{\text{Aufgabe}.}$ = 2.6, SD_{Aufgabe} = 0.98) und die Zeit zu managen (M_{Testat} = 2.9, SD_{Testat} = 0.66 vs. $M_{\text{Aufgabe}.}$ = 2.3, SD_{Aufgabe} = 1.00).

4.3 Gestaltungsspielräume zum selbstregulierten Lernen

Zu den Dimensionen selbstregulierten Lernens lagen 63 vollständige Datensätze vor. Für paarweise Vergleiche wären 84 Personen notwendig gewesen, daher wurde auf eine inferenzstatistische Auswertung verzichtet. Die Studierenden nahmen die ihnen gebotenen Handlungsspielräume ziemlich bis sehr wahr für das Festlegen der Inhalte (M = 4.6, SD = 0.58), des Orts (M = 4.6, SD = 0.91), der Zeit (M = 4.4 SD = 0.94), der Methoden (M = 4.2, SD = 0.82), der Quellen (M = 4.0, SD = 0.94) und der Ziele ihres Lernens (M = 3.8, SD = 0.89). Weniger sahen sie ihren Handlungsspielraum für das Wählen der Lernpartner:in (M = 3.0, SD = 1.40) und der Einschätzung des eigenen Lernens (M = 2.8, SD = 1.0).

Ziemlich bis sehr nutzten die Studierenden die Möglichkeiten Ort (M = 4.6, SD = 0.69), Zeit (M = 4.5, SD = 0.91), Inhalte (M = 4.2, SD = 0.95), Methoden (M = 4.1, SD = 0.76) und Quellen des Lernens (M = 4.0, SD = 0.92) auszuwählen. Weniger wurde der Handlungsspielraum im Hinblick auf Lernziele (M = 3.5, SD = 1.03), der Lernpartner:in (M = 3.3, SD = 1.53) und insbesondere der Lerneinschätzung (M = 2.7, SD = 1.05) ausgeschöpft.

Als mittelmäßig bis sehr förderlich für den Lernprozess nahmen die Befragten das Festlegen der Lerninhalte wahr ($M = 4.3$, $SD = 0.69$). Mittelmäßig bis ziemlich förderlich bewerteten sie es, Quellen ($M = 4.0$, $SD = 0.87$), Zeit ($M = 4.0$, $SD = 1.17$), Methoden ($M = 3.8$, $SD = 0.90$), Ort ($M = 3.7$, $SD = 1.26$) und Ziele ($M = 3.5$, $SD = 0.85$) festzulegen sowie das eigene Lernen einzuschätzen ($M = 3.0$, $SD = 1.18$). Dagegen wurde das Auswählen der Lernpartner:in als wenig bis mittelmäßig hilfreich empfunden ($M = 2.6$, $SD = 1.40$).

Mittelmäßig bis ziemlich herausfordernd für den eigenen Lernprozess empfanden die Studierenden das eigene Lernen einzuschätzen ($M = 3.5$, $SD = 1.13$), Inhalte festzulegen ($M = 3.2$, $SD = 0.99$) und Quellen auszuwählen ($M = 3.0$, $SD = 0.91$). Nicht bis mittelmäßig herausfordernd wurden das Auswählen von Ort ($M = 1.6$, $SD = 0.78$), Partner:innen ($M = 1.6$, $SD = 1.12$), Zeit ($M = 2.5$, $SD = 1.24$), Methoden ($M = 2.7$, $SD = 0.91$) und Zielen ($M = 2.9$, $SD = 1.00$) empfunden.

5. Diskussion

Zunächst werden die Ergebnisse der Evaluation zu den einzelnen Fragestellungen zusammengefasst und eingeordnet. Dann wird dargestellt, wie die Prüfungstheke anhand der Ergebnisse angepasst wurde und diskutiert, inwiefern generative künstliche Intelligenz (KI) die Weiterentwicklung der Prüfungstheke beeinflusst.

Wie wird die Prüfungstheke durch die Studierenden angenommen? Die meisten eingeschriebenen Studierenden absolvierten die Teilleistungen der Vorlesung Ingenieurpsychologie erfolgreich und probierten neue Prüfungsformate deutlich häufiger aus als zwingend erforderlich gewesen wäre. Insgesamt sprechen die Ergebnisse für ein hohes Maß an Akzeptanz der Studierenden für die Prüfungstheke. Dieser Befund wird durch die hohe Zufriedenheit untermauert, wobei sich die Zufriedenheit mit der Erfahrung der Prüfungstheke eher verstetigt und nicht allein auf einem kurzfristigen Neuheitseffekt (Clark, 1983; Keller & Suzuki, 2004) zu beruhen scheint.

Wie sind die Portfolioaufgaben als Prüfungsleistungen im Vergleich zu traditionellen Testaten zu bewerten? Die Retest-Reliabilität der Portfolioaufgaben erwies sich als mangelhaft, dagegen zeigte sich die der Testate als moderat bis gut. Dabei ist zu berücksichtigen, dass verschiedenartige Portfolioaufgaben zu unterschiedlichen Zeitpunkten bearbeitet und die homogenen Testate am selben Tag geschrieben wurden. Grundsätzlich ist fraglich, ob stabile Bewertungsergebnisse bei Portfolioaufgaben, die einen kontinuierlichen Kompetenzerwerb begleiten, überhaupt erstrebenswert sind. Die Ergebnisse der Portfolioaufgaben und Testate hingen nur gering bis mäßig zusammen. Als Prüfungen derselben Lehrveranstaltung mit identischen Lehrinhalten wäre ein höherer Zusammenhang zu erwarten gewesen. Dabei ist zu berücksichtigen, dass sowohl für die Portfolioaufgaben als auch für die Testate hohe Punktwerte vergeben wurden. Wie in der Psychologie üblich (Gaens & Müller-

Benedict, 2017), wurde nicht das ganze Spektrum der verfügbaren Punkteskala ausgeschöpft. Diese eingeschränkte Streuung verringert die potenzielle Höhe der Korrelation zwischen Testat- und Aufgabenpunktwerten. Eine geringe Reliabilität eines oder beider Prüfungsformate kann ebenfalls zu der geringen Übereinstimmung beider Prüfungsformate beigetragen haben (Bühner, 2021). Auch ist offen, ob alternative Portfolioaufgaben und Testate summativ verschiedene Kompetenzen messen oder ob mit den verschiedenen Formaten unterschiedliche kognitive Prozesse angeregt werden. Ohne Außenkriterium, wie der Leistung in einer authentischen Aufgabe, lässt sich nicht beurteilen, wie gut welche Kompetenzen tatsächlich erworben und überprüft wurden. Der Vergleich der Lernstrategien legt nahe, dass mit Portfolioaufgaben eine tiefere Verarbeitung der Lerninhalte einhergeht, wobei offen ist, ob die Aufgaben die Strategien beeinflussen oder Personen, die tiefere Verarbeitung bevorzugen, eher Portfolioaufgaben wählen. Im Mittel bearbeiteten die Studierenden die Portfolioaufgaben zu allen drei Zeitpunkten 13 bis 15 Stunden. Da pro Themenblock 25 Stunden geplant waren, stellt die Prüfungstheke aus Studierendensicht seine sehr ökonomische Prüfung dar. Für die Testate zeichnet sich ab, dass die Lernbelastung gegen Ende ansteigt, wobei in weiteren Evaluationen die Beteiligung der Testatschreibenden deutlich erhöht werden sollte, um abschließend zu beurteilen, inwiefern die Prüfungstheke nachhaltiges, verteiltes Lernen fördert. Trotz der unterschiedlichen Zuverlässigkeit und des geringen Zusammenhangs zwischen den Prüfungsleistungen zeigten sich zwischen neuartigen Aufgaben und traditionellem Testat keine grundsätzlichen Unterschiede im Bewertungsniveau. Insgesamt handelt es sich bei der Prüfungstheke um ein Konzept, dass bei großen Studierendengruppen individualisierte und zumutbare Prüfungen erlaubt. Die Gültigkeit der erfassten Kompetenzen kann noch nicht abschließend eingeschätzt werden.

Die Studierenden nehmen die sich mit der Prüfungstheke eröffneten Freiräume wahr und nutzen diese. Für das Auswählen eigener Ziele und Inhalte sehen die Studierenden einen größeren Spielraum als sie ihn nutzen. Die Option, das eigene Lernen einzuschätzen, wird verglichen mit den anderen Dimensionen selbstregulierten Lernens deutlich weniger wahrgenommen und genutzt. Gleichzeitig sehen die Studierenden hierbei die größte Herausforderung. Das mag daran liegen, dass in dem Studiengang bislang reflexive Praxis wenig eingeübt wird und die Studierenden dazu mehr Unterstützung benötigen. Gleichzeitig sehen die Befragten den größten Entwicklungsbedarf der Prüfungstheke in lernförderlichem Feedback zu den Portfolioaufgaben. Sie wünschen sich elaborierte, zeitnahe Rückmeldung, um aus Fehlern lernen zu können. Daher wird in der überarbeiteten Prüfungstheke die Bewertungslast der Dozierenden gemindert, indem weniger Aufgaben nach weniger Kriterien mit höchstens drei Abstufungen bewertet werden. Die Studierenden erhalten die Bewertungskriterien als ausformulierte Rubrics (Kriterienraster), um formativ ihren Lernstand einschätzen zu können (Panadero et al., 2023).

Welche Rolle spielt die Verfügbarkeit generativer KI-Tools wie Chat-GPT für die Weiterentwicklung der Prüfungstheke? Einerseits wird befürchtet, dass es mehr Täuschungsversuche geben wird und diese schwieriger zu identifizieren seien (Fleischmann, 2023). Andererseits kann KI sinnvoll als eine Art Lernbegleiter für Studierende eingesetzt werden und Lehrende entlasten.

An der TU Dresden, an der die Prüfungstheke umgesetzt wurde, gibt es (noch) keine einheitlichen Regelungen zum Umgang mit generativer KI in der Lehre. Wir müssen davon ausgehen, dass generative KI-Tools wahrscheinlich integriert in andere Tools wie BING oder Copilot von vielen Studierenden bereits genutzt werden (Wietz, 2024). Je nach Prüfungsformat ist damit die Nutzung mehr oder weniger wahrscheinlich. Die Prüfungstheke setzt an einer Stelle an, an der die KI noch nicht inhaltlich unterstützen kann. Wenn ich die KI (Microsoft Copilot) frage, was für eine Prüfungsform ich wählen soll, erhalte ich den metakognitiven Tipp: »Letztendlich hängt die Wahl der Prüfungsform von deinen persönlichen Vorlieben und Lernzielen ab. Überlege, welche Methode am besten zu deinem individuellen Lernstil passt und wie du das Thema am effektivsten vertiefen möchtest. Viel Erfolg!«. Damit sind Täuschungsversuche zwar nicht ausgeschlossen, aber Studierende müssen sich dennoch im Vorfeld mit dem Inhalt und der damit verbundenen Wahl der Prüfungsleistung beschäftigen.

Damit ist für uns ein Weg angedeutet, wie generative KI als Lernbegleiter fungieren kann. Selbst wenn es KI-Tools schaffen, z.B. sinnvolle Concept-Maps zu erstellen, stellen diese zunächst nur einen Ausgangspunkt dar. Von dem aus müssen sich Lernende intensiv mit den Inhalten auseinandersetzen, wie beim Lernen an (auch unvollständigen) Beispielen (Renkl et al., 2004). Als weitere Prüfungsform kann im Dialog mit generativer KI ein Problemfeld umrissen oder eine Fragestellung bearbeitetet werden. Die Dokumentation dieser Auseinandersetzung (mit Reflexion der Vorgehensweise) entspräche dann einer schriftlichen Arbeit.

In der Vision einer neuen Prüfungskultur (Budde et al., 2024) spielt KI ebenfalls eine Rolle. So wird beispielsweise ein Szenario beschrieben, in dem die Begleitung durch Lehrende mit Unterstützung durch KI-Systeme kombiniert wird. Auch wenn hierzu noch keine konkreten Konzepte für die Weiterentwicklung der Prüfungstheke vorliegen, so wäre dies eine Möglichkeit, um formatives Assessment stärker in den Lernprozess über das Semester zu implementieren.

Weiterhin wird in der Vision der neuen Prüfungskultur Flexibilität als ein Kernelement postuliert: »Die Heterogenität und Diversität der Studierenden wird berücksichtigt und gleichzeitig Fairness und Chancengleichheit gewährleistet.« (ebd., 11). Die Prüfungstheke nähert sich diesem Gedanken an, indem die unterschiedlichen Prüfungsformate nicht nur individuelle Kompetenzentwicklung begünstigen, sondern auch eine flexible Verteilung der Prüfungslast zulassen.

Literaturverzeichnis

Albrecht, C., Schmidt, J., & Jantos, A. (2023). Die Prüfungstheke als Prüfungsstrategie der Zukunft, *Zeitschrift für Hochschulentwicklung*, *18*(3), 213–239. https://doi.org/10.21240/zfhe/18-03/11

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice, 24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*(6), 445–457. https://doi.org/10.1016/S0883-0355(99)00014-2

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. Journal of Cognition, 2(1). https://doi.org/10.5334/joc.72

Budde, J., Tobor, J., & Eichhorn, J. (2024). *Vision einer neuen Prüfungskultur*. https://hochschulforumdigitalisierung.de/wp-content/uploads/2024/01/HFD_Diskussionspapier_28_Vision-einer_neuen_Pruefungskultur_final.pdf?wt_zmc = nl.int.zonaudev.112331552451_447560032380.nl_ref.

Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion* (4., korrigierte und erweiterte Auflage). Pearson.

Caldwell, A. R. (2022). Exploring equivalence testing with the updated TOSTER R package. *PsyArXiv*. https://doi.org/10.31234/osf.io/ty8de \

Champely, S. (2020). Pwr: Basic Functions for Power Analysis.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459. https://doi.org/10.3102/00346543053004445

Dyrna, J. (2021). Selbstgesteuertes Lernen, Begriffsbestimmung und Operationalisierung. In J. Dyrna, J. Riedel, S. Schulze-Achatz & T. Köhler (Hg.), *Selbstgesteuertes Lernen in der beruflichen Weiterbildung* (S. 65–85). Waxmann.

Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests Test Review Form and Notes for Reviewers, Version 4.2.6.*

Fleischmann, A. (2023). ChatGPT in der Hochschullehre. Wie künstliche Intelligenz uns unterstützen und herausfordern wird. In B. Berendt (Hg.), *Neues Handbuch Hochschullehre* (S. 13–56). DUZ. https://www.nhhl-bibliothek.de

Frank, A. (o.J.). *Bielefelder Lernzielorientierte Evaluation (BiLOE)* – Universität Bielefeld. https://www.uni-bielefeld.de/einrichtungen/zll/hdle/evaluation/biloe/

Friedrich, H. F., & Mandl, H. (1997). Analyse und Förderung selbstgesteuerten Lernens. In F. E. Weinert & H. Mandl (Hg.), *Themenbereich D: Praxisgebiete, Serie I Pädagogische Psychologie: Bd. Psychologie der Erwachsenenbildung* (S. 237–296). Hogrefe.

Gaens, T., & Müller-Benedict, V. (2017). Die langfristige Entwicklung des Notenniveaus und ihre Erklärung. In V. Müller-Benedict & G. Grözinger (Hg.), *Noten an Deutschlands Hochschulen: Analysen zur Vergleichbarkeit von Examensnoten 1960 bis 2013* (S. 17–78). Springer Fachmedien. https://doi.org/10.1007/978-3-658-15801-9_1

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

Keller, J., & Suzuki, K. (2004). Learner motivation and E-learning design: A multinationally validated process. *Journal of Educational Media, 29*(3), 229–239. https://doi.org/10.1080/1358165042000283084

Klingsieck, K. B. (2018). Kurz und knapp – die Kurzskala des Fragebogens »Lernstrategien im Studium« (LIST). *Zeitschrift für Pädagogische Psychologie*, *32*(4), 249–259. https://doi.org/10.1024/1010-0652/a000230

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4), 355–362. https://doi.org/10.1177/1948550617697177

Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review, 35*(4), 113. https://doi.org/10.1007/s10648-023-09823-4

Perrella, A., Koenig, J., Kwon, H., Nastos, S., & Rangachari, P. K. (2015). On being examined: Do students and faculty agree? *Advances in Physiology Education, 39*(4), 320–326. https://doi.org/10.1152/advan.00078.2015

R Core Team. (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Reinmann, G. (2022). Prüfung oder Assessment an Hochschulen? Thesen für einen Wandel der Prüfungskultur. In J. Gerick, A. Sommer & G. Zimmermann (Hg.), *Kompetent Prüfungen gestalten* (2. überarb. u. erw. Aufl., S. 22–36). Waxmann. https://elibrary.utb.de/doi/book/10.36198/9783838558592

Renkl, A., Schworm, S., Hilbert, T. S. (2004). Lernen aus Lösungsbeispielen. Eine effektive, aber kaum genutzte Möglichkeit, Unterricht zu gestalten. In J. Doll (Hg.), *Bildungsqualität von Schule. Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (S. 77–92). Waxmann.

Revelle, W. (2024). Psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University.

Schaper, N. (2021). Prüfen in der Hochschullehre. In R. Kordts-Freudinger, N. Schaper, A. Scholkmann & B. Szczyrba (Hg.), *Handbuch Hochschuldidaktik* (S. 87–101). wbv Publikation.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry, 31*(2), 245–261. https://doi.org/10.1111/j.1465-7295.1993.tb00881.x

Wietz, M., Schrimph, L., & Rohr, N. (2024). KI oder Kreide im Hörsaal – so digital sind Deutschlands Hochschulen. *Bitkom Research*. https://www.bitkom.org/Presse/Presseinformation/So-digital-sind-Deutschlands-Hochschulen