



KI-gestützte Lernsimulation für die sprachtherapeutische Hochschullehre: Konzeption und formative Expertinnenevaluation eines Trainings zum Dynamic Assessment grammatischer Fähigkeiten*

AI-Supported Learning Simulation for Speech-Language Therapy Education: Design and Formative Expert Evaluation of a Training Tool for Dynamic Assessment of Grammatical Abilities

Johannes Winterstein, Maximilian Hamann

Zusammenfassung

Hintergrund: Diagnostische Kompetenzen, insbesondere das Dynamic Assessment (DA) grammatischer Fähigkeiten, lassen sich in der sprachtherapeutischen Hochschullehre nur schwer systematisch und wiederholbar trainieren.
Fragestellung/Ziele: Der Beitrag stellt eine generative KI-gestützte Simulationsumgebung zum Training des DA vor und prüft explorativ deren Umsetzbarkeit und didaktischen Mehrwert.
Methodik: Die Simulation kombiniert eine deterministische Antwortmatrix mit einem Sprachmodell und bildet eine triadische Lehrsituation ab, in der Studierende ein simuliertes Kind diagnostisch untersuchen und dabei von einem KI-Meta-Trainer begleitet werden. Sie wurde im ersten Zyklus eines Design-Based-Research-Prozesses durch eine formative Expertinnenevaluation (n = 3) mit Fragebögen und einem leitfadengestützten Reflexionsgespräch eingeschätzt.
Ergebnisse: Die Expertinnen bewerten das didaktische Konzept durchgehend hoch, mit Spitzenwerten beim Clinical Reasoning sowie der Curriculum-Integration und benennen eine studierendenseitige Vorbereitungs-Hürde. Den Hauptüberarbeitungsbedarf verorten sie in der Benutzeroberfläche. Die inhaltlichen Rückmeldungen des Meta-Trainers werden geschätzt; sein Eingreifen nach jedem Item wird einer durchgehenden Interaktion mit dem Kind allerdings nicht vorgezogen.
Schlussfolgerungen: Die Umgebung erscheint fachlich anschlussfähig und curricular als Vertiefungsinstrument geeignet, wobei das Design keine Aussagen zur Lernwirksamkeit erlaubt. Die vier Gestaltungsprinzipien werden als Hypothesen für die Übertragung auf benachbarte Diagnostik-Domänen formuliert.

Schlüsselwörter

Dynamic Assessment, diagnostische Kompetenz, KI-gestützte Simulation, Sprachtherapie, Hochschullehre, Design-Based Research

* Dieser Beitrag hat das double-blind Peer-Review-Verfahren durchlaufen.

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

Abstract

Background: Diagnostic competences, in particular the Dynamic Assessment (DA) of grammatical abilities, are difficult to train systematically and repeatably in speech-language therapy education.
Aims: This paper presents a generative-AI simulation environment for training DA and explores its feasibility and didactic value.
Methods: The simulation combines a deterministic answer matrix with a language model and enacts a triadic teaching situation in which students diagnostically assess a simulated child while being supported by an AI meta-trainer. It was evaluated in the first cycle of a design-based-research process through a formative expert evaluation (n = 3) using questionnaires and a guideline-based reflective interview.
Results: Experts rated the didactic concept consistently high, with peak ratings for clinical reasoning and curriculum integration and identified a student-side preparation prerequisite. They located the main revision need in the user interface. The meta-trainer's feedback was valued, but its timing after every item was not preferred over an uninterrupted interaction with the child.
Conclusions: The environment appears professionally sound and suitable as a curricular consolidation tool. The design does not permit claims about learning effectiveness. The four design principles are formulated as hypotheses for transfer to adjacent diagnostic domains.

Keywords

dynamic assessment, diagnostic competence, AI-based simulation, speech-language therapy, higher education, clinical reasoning

1 Einleitung

Die Entwicklung diagnostischer Kompetenzen stellt ein zentrales Lernziel in der Ausbildung von Sprachtherapeut:innen dar, denn die Qualität therapeutischer Entscheidungen hängt unmittelbar von einer tragfähigen Diagnostik ab. Professionelles diagnostisches Handeln beruht dabei nicht allein auf Fachwissen, sondern auf dessen strukturierter, fallbezogener und methodisch kontrollierter Anwendung (Beushausen & Walther, 2010; Dudding & Pfeiffer, 2018; Spreer, 2018). Studierende müssen lernen, sprachliche Auffälligkeiten zu analysieren, Befunde zu integrieren, Hypothesen zu prüfen und daraus begründete diagnostische sowie therapeutische Entscheidungen abzuleiten. Dieses diagnostische Clinical Reasoning ist zentral für die Identifikation individueller Therapiebedarfe und die Auswahl evidenzbasierter Interventionen (Diepeveen et al., 2020; Hoben et al., 2007; Saccone et al., 2025). Besondere Anforderungen stellt das Dynamic Assessment (DA), das neben dem aktuellen Leistungsstand auch das sprachliche Lernpotenzial eines Kindes und dessen Reaktion auf gestufte Hilfen einbezieht und damit hilft, die Ursachen sprachlicher Schwierigkeiten differenzierter einzuschätzen (Ehlert, 2021a). Gerade solche dynamisch-diagnostischen Prozesse sind in der Hochschullehre jedoch schwer systematisch zu trainieren, da reale Diagnostiksituationen zeitlich begrenzt, organisatorisch aufwendig, ethisch sensibel und didaktisch nur eingeschränkt steuerbar sind (Heitzmann et al., 2019; Nagdee et al., 2022).

KI-gestützte Lernsimulationsumgebungen eröffnen hier neue Möglichkeiten, indem sie adaptive Fallsituationen bereitstellen, diagnostische Entscheidungen einfordern und individualisiertes Feedback geben (Chernikova et al., 2020a; Hill et al., 2021). Der Forschungsstand ist dabei jedoch ungleich verteilt. Für das simulationsbasierte Training diagnostischer Kompetenzen liegt mit der Metaanalyse von Chernikova et al. (2020b) zwar eine breite Evidenzbasis vor, diese stützt sich allerdings auf die Medizin- und Lehrkräftebildung. Für die Sprachtherapie ist die Befundlage nicht nur schmaler, sondern auch heterogener in ihrer Aussagekraft. Am belastbarsten ist die randomisierte kontrollierte Studie von Hill et al. (2021), wonach simulationsbasiertes Lernen einen Teil der Praktikumszeit ersetzen kann. Ergänzend belegt Carter (2019) Leistungszuwächse durch computerbasierte Simulationen. Die übrigen einschlägigen Arbeiten liefern dagegen keine Wirksamkeitsnachweise. Nagdee et al. (2022) ordnen die Formate in einem Scoping Review systematisch ein, während Hewat et al. (2020) ihr Entwicklungsframework lediglich aus der Reflexion über ein einzelnes fünftägiges Programm ableiten. Für KI-basierte Simulationen fehlt eine solche Evidenz in der Sprachtherapie bislang gänzlich, da sich die vorliegenden Anwendungen überwiegend auf Therapie- und Dokumentationsaufgaben richten und nicht auf die Ausbildung (Georgiou, 2025; Suh et al., 2024).

Für das Dynamic Assessment selbst fehlt dieser Anschluss. Eine orientierende Recherche in PubMed, ERIC und Google Scholar, die keinen Anspruch auf Systematik erhebt, verband Such-

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

begriffe zum Dynamic Assessment mit solchen zu Simulation, Hochschullehre und generativer KI. Sie fand Fortbildungsangebote zum DA für bereits berufstätige Fachpersonen sowie Simulationsstudien zu anderen sprachtherapeutischen Gegenständen, etwa zur Dysphagie oder zur unterstützten Kommunikation. Eine simulationsbasierte Trainingsumgebung, die das DA selbst zum Lerngegenstand macht, war darunter nicht. Diese Lücke ist der Ausgangspunkt des vorliegenden Beitrags.
Daraus ergeben sich zwei aufeinander bezogene Fragestellungen. Die erste ist eine Gestaltungsfrage: Wie wurde die vorliegende KI-Simulationsumgebung didaktisch und technisch gestaltet, damit Studierende das DA grammatischer Fähigkeiten systematisch und wiederholbar trainieren können (Kapitel 2 und3)? Die zweite ist eine Evaluationsfrage und wird in Kapitel 4 beantwortet: Wie beurteilen sprachtherapeutische Expertinnen die Umsetzbarkeit und den didaktischen Mehrwert dieser Simulation?

2 Theoretischer Rahmen

2.1 Diagnostische Kompetenz in der Sprachtherapie

Diagnostische Kompetenzen zählen zu den zentralen Inhalten in den Curricula der akademischen Sprachtherapie und werden auch in den Empfehlungen der gesetzlichen Krankenversicherungen zu den Ausbildungsinhalten ausdrücklich als qualifikationsrelevant ausgewiesen (American Speech-Language-Hearing Association [ASHA], 2016; Beushausen & Walther, 2010). Ihr konzeptueller Kern ist das diagnostische Clinical Reasoning, also der fachliche Denk- und Entscheidungsprozess, in dem sprachtherapeutische Befunde systematisch erhoben, kritisch bewertet und zu einer nachvollziehbar begründeten diagnostischen Entscheidung zusammengeführt werden. Gerade dieser Kern ist jedoch theoretisch wie didaktisch bislang nur begrenzt vereinheitlicht. Die konzeptionelle Unschärfe erschwert es, diagnostisches Reasoning in der Hochschullehre systematisch zu vermitteln und zu bewerten; sichtbar wird dies unter anderem daran, dass Studierende in diagnostischen Fallsituationen sehr unterschiedliche Strategien der Hypothesenbildung, Befundintegration und Begründung nutzen (Saccone et al., 2025; Gama et al., 2025).

Einen tragfähigen Ordnungsrahmen für diese Unschärfe bietet die interdisziplinäre Bildungsforschung. Fischer et al. (2014) modellieren wissenschaftliches Denken und Handeln anhand acht epistemischer Aktivitäten, die von der Problemidentifikation und Hypothesenbildung bis zum Schlussfolgern und Kommunizieren reichen. Heitzmann et al. (2019) bauen auf diesem Modell auf. Sie definieren diagnostische Kompetenzen als individuelle Dispositionen, die es ermöglichen, vorhandenes Wissen nach professionellen Standards in diagnostischen Aktivitäten anzuwenden, um Daten zu sammeln, zu interpretieren und daraus qualitativ hochwertige Entscheidungen abzuleiten. Bauer et al. (2020) zeigen schließlich, dass sich diese Aktivitäten je nach Domäne zu spezifischen diagnostischen Praktiken (diagnostic practices) ausdifferenzieren. Für die Sprachtherapie lässt sich diagnostische Kompetenz damit als domänenspezifische Disposition fassen, die Fachwissen, Methodenkompetenz und fallbezogenes Reasoning verbindet.

Diese Definition wird didaktisch auswertbar, wenn man drei Ebenen diagnostischen Handelns unterscheidet. Die methodische Durchführung bildet als Performanzebene das beobachtbare diagnostische Handeln in der konkreten Testsituation ab. Das diagnostische Urteil führt als Befundintegration einzelne Beobachtungen, Produktionsmuster und Reaktionen auf Hilfen zusammen. Die diagnostische Argumentation begründet schließlich dieses Urteil. Für diese dritte Ebene differenzieren Bauer et al. (2022) drei Facetten, die angehende Diagnostiker:innen nur selten gemeinsam realisieren: Justification als evidenzbezogene Begründung des diagnostischen Urteils, Disconfirmation als Prüfung plausibler Alternativerklärungen für die beobachteten Leistungen und Transparency als Offenlegung des Prozesses, durch den diagnostische Evidenz erzeugt und gewichtet wurde. Für die Diagnostik grammatischer Fähigkeiten konkretisieren sich die Ebenen diagnostischen Handelns in der Formenanalyse morphosyntaktischer Strukturen (Durchführung), der Befundintegration über mehrere Datenquellen (Urteil) und der begründeten Ableitung therapeutischer Entscheidungen (Argumentation); sie bilden damit die Lernziele der in 3.1 beschriebenen Simulation.

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

2.2 Dynamic Assessment

Dynamic Assessment (DA) bezeichnet eine diagnostische Vorgehensweise, die sprachliche Leistungen nicht statisch erfasst, sondern im Zusammenhang mit Lernunterstützung, Reaktion auf Hilfen und Veränderbarkeit betrachtet (Delage et al., 2021; Ehlert, 2021a). Im Anschluss an Vygotskijs Konzept der Zone der nächsten Entwicklung (Vygotsky, 1978) rückt damit das sprachliche Lernpotenzial in den Mittelpunkt: Entscheidend ist nicht allein der aktuelle Leistungsstand, sondern welche Leistungen unter systematischer Unterstützung erreichbar werden und wie flexibel ein Kind auf Instruktion, Rückmeldung und Modellierung reagiert.

Für die Sprachtherapie ist DA besonders relevant, weil diagnostisch häufig zu klären ist, ob beobachtete Schwierigkeiten auf mangelnde Erfahrung, mehrsprachige Erwerbsbedingungen oder eine Sprachentwicklungsstörung (SES; Konsensusdefinition nach Kauschke et al., 2023) zurückzuführen sind. Dahinter steht eine differenzialdiagnostische Logik: Kinder mit ungünstigen Lernbedingungen oder mehrsprachigem Hintergrund reagieren bei intaktem Sprachlernpotenzial auf strukturierte Unterstützung eher als Kinder mit einer SES. Gerade in mehrsprachigen Kontexten gilt DA daher als vielversprechend, weil nicht primär das in einer Testsprache erworbene Wissen, sondern die Lernreaktion auf Unterstützung beurteilt wird (Ehlert, 2021b). Empirisch stützen Schwob et al. (2025) dies: Sie verglichen an 49 ein- und mehrsprachigen Kindern von 5;0 bis 7;11 Jahren zwei DA-Situationen, ein eigenständig bearbeitetes Computerspiel und ein interaktives Vorlesesetting mit gestuften Hilfen. Die beste Klassifikation gelang mit überwiegend rezeptiven Aufgaben aus dem interaktiven Setting (bis zu 100 % Sensitivität und 96 % Spezifität), ohne dass der Sprachstatus der Kinder darauf Einfluss hatte. Gerade die gestufte Hilfengebung, also das Prinzip des Graduated Prompting, trug die diagnostische Trennschärfe. Diese Logik greift die vorliegende Simulation im DA-Format Graduated Prompting (GP) auf. Gefragt ist nicht nur, ob eine Zielantwort erreicht wird, sondern auf welcher Hilfestufe das gelingt und wie sich die Leistung im Verlauf der Unterstützung verändert, sodass Veränderbarkeit als Prozessmerkmal und nicht erst als Ergebnis sichtbar wird. Ergänzend dazu schätzt die untersuchende Person fortlaufend die Modifiability des Kindes ein, also dessen individuelle Veränderbarkeit während der Testsituation. Gemeint ist, wie gut das Kind auf Hilfen anspricht, Lernangebote aufnimmt und seine Leistung unter Anleitung anpasst. Den unterstützenden Rahmen dafür beschreibt die DA-Tradition als vermittelte Lernerfahrung (Mediated Learning Experience) (Delage et al., 2021; Ehlert, 2021a).

Für die Erfassung grammatischer Fähigkeiten ist DA besonders anschlussfähig, weil morphosyntaktische Leistungen im Prozess ihrer Elizitierung unmittelbar beobachtbar werden. Dadurch kann die untersuchende Person die drei in 2.1 beschriebenen Ebenen diagnostischen Handelns (methodische Durchführung, Befundintegration und diagnostische Argumentation) innerhalb einer Aufgabe aufeinander beziehen. Für die Morphosyntax belegen das Hadjadj et al. (2022), die 79 französischsprachige Kinder von 5;1 bis 11;9 Jahren mit gestuften Hilfen vom metakognitiven Hinweis bis zur Imitationsaufforderung prüften. Die Mehrsprachigkeit beeinflusste die Werte nicht, die diagnostische Gruppe dagegen deutlich. Trennscharf waren allerdings nur komplexe Strukturen mit Bewegung oder Einbettung und hinreichende Trefferquoten ergaben sich erst bei Kindern mit unabhängig belegtem morphosyntaktischem Defizit, was die Wahl von Zielstrukturen mit Verbbewegung stützt. Trotz dieses diagnostischen Potentials ist DA in der sprachtherapeutischen Praxis noch wenig etabliert (Delage et al., 2021; Ehlert, 2021a). Aus dieser Spannung erwächst der Hochschullehre die Aufgabe, DA früh zu vermitteln und wiederholbare Trainingsgelegenheiten für Hilfengebung, Formenanalyse, Hypothesenbildung und diagnostische Begründung zu schaffen.

2.3 KI-gestützte Simulationen in der Hochschullehre

KI-gestützte Simulationen bieten die Möglichkeit, diagnostische Kompetenzen bereits im Studium gezielt zu fördern, indem sie diagnostische Situationen als Approximations of Practice, also handlungsnahe Annäherungen an die spätere Berufspraxis darstellen (Chernikova et al., 2020a). Sie schaffen kontrollierte Lernumgebungen, in denen Studierende diagnostische Situationen bearbeiten, Entscheidungen treffen, Rückmeldung erhalten und ihr Vorgehen reflektieren können, ohne reale Patient:innen einzubeziehen.
Ein zentraler Wirkmechanismus solcher Simulationen sind authentische Fälle, denn erst die Anwendung theoretischen Wissens auf konkrete klinische, therapeutische oder pädagogische

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

Problemstellungen trägt zum Aufbau professioneller Expertise bei (Fischer et al., 2022). Die wiederholte Fallbearbeitung fördert dabei den Wissenstransfer und den Aufbau handlungsnaher Entscheidungsschemata (Sailer et al., 2023) und erlaubt es zugleich, die acht epistemischen Aktivitäten nach Fischer et al. (2014) einzuüben, die Heitzmann et al. (2019) für diagnostische Situationen ausdifferenzieren.

Der spezifische Mehrwert KI-gestützter Formate zeigt sich im Vergleich mit etablierten fallbasierten Formaten. Papiervignetten und Videofälle sind standardisiert, reagieren aber nicht; Rollenspiele mit Peers oder Schauspielpatient:innen reagieren, lassen sich aber weder standardisieren noch beliebig wiederholen. KI-gestützte Formate verbinden beides und ergänzen eine adaptive, prozessorientierte Rückmeldung. Verfahren des Natural Language Processing (NLP), also der automatisierten Verarbeitung natürlicher Sprache, analysieren dafür freitextliche Begründungen, Hypothesen und Reflexionen und erkennen darin inhaltliche Merkmale. Auf dieser Analyse setzt eine Rückmeldekomponente auf, die zeitnah und individualisiert antwortet (Bauer et al., 2023; Sailer et al., 2023). NLP erzeugt die Rückmeldung also nicht selbst, sondern liefert ihre Grundlage. Rückmeldung zielt damit nicht nur auf die Ergebnisrichtigkeit, sondern auf das diagnostische Reasoning. Bewertet wird, ob relevante Informationen berücksichtigt, Hypothesen evidenzbasiert begründet, Alternativen geprüft und Entscheidungswege transparent gemacht werden, also ebendie Facetten, die Bauer et al. (2022) als Justification, Disconfirmation und Transparency beschreiben.

Drei pragmatische Bedingungen machen simulationsbasiertes Training darüber hinaus für die Hochschullehre tragfähig. Erstens können Studierende Fehler machen und Entscheidungen revidieren, ohne Patient:innen zu gefährden, sodass Fehler diagnostisch lehrreich statt klinisch folgenreich werden (Hewat et al., 2020). Zweitens erlauben wiederholbare Durchgänge den Aufbau diagnostischer Routinen, die in der zeitlich begrenzten Praxis schwer erreichbar sind. Drittens ermöglichen digitale Simulationen größeren Studierendengruppen Zugang zu vergleichbaren Fallszenarien bei individualisiertem Feedback (Kiesewetter et al., 2022; Sailer et al., 2023). Für die sprachtherapeutische Hochschulbildung entsteht so ein vorbereitender Trainingsraum, der die drei Ebenen diagnostischen Handelns (Durchführung, Befundintegration und Argumentation) zusammenführt. Die hier vorgestellte Simulation überträgt dieses Modell erstmals auf das DA grammatischer Fähigkeiten, für das bislang kein etabliertes Trainingsformat vorliegt.

2.4 SHARP-Framework als didaktischer Strukturrahmen

Der vorliegende Beitrag verwendet SHARP (Simulation-Based Learning in Higher Education: Advancing Research on Process Diagnostics and Personalised Interventions; Bauer et al., 2025), ein Rahmenmodell für personalisiertes simulationsbasiertes Lernen in der Hochschulbildung. Es dient hier heuristisch als Designgerüst, nicht als hypothesentestbares Modell, weshalb die Expertinnenstudie in Kapitel 4 keine SHARP-spezifischen Annahmen prüft, sondern das daraus abgeleitete Design bewertet. Der Kerngedanke des Frameworks besteht darin, die zentralen Designentscheidungen einer Simulation durch fünf Leitfragen explizit zu machen: Warum wird personalisiert, was wird personalisiert, wer personalisiert, wie wird personalisiert und wann wird personalisiert?

Abbildung 1 überträgt das Rahmenmodell auf die vorliegende Simulation. Entlang des Lernprozesses gelesen, beantwortet sie die fünf Leitfragen wie folgt. Die Lernvoraussetzungen erfasst ein mehrdimensionaler Vortest, denn Studierende unterscheiden sich erheblich in ihrer Vertrautheit mit MuSE-Pro (Berg, 2015; vgl. 3.2), mit dem DA und mit der GP-Hilfehierarchie (warum personalisiert wird). Die Lernunterstützung leistet der Meta-Trainer, abgestuft über drei Support-Level (was personalisiert wird). Die Lerninhalte repräsentieren die drei Fallvignetten. Die Personalisierung erfolgt fallübergreifend, jeweils nach dem Debriefing eines Falls (wann personalisiert wird), wobei das System das Support-Level setzt und die Studierenden die Fallprogression wählen (wer und wie personalisiert wird). Die Lernergebnisse schließlich folgen den acht epistemischen Aktivitäten nach Fischer et al. (2014) in ihrer diagnostischen Ausdifferenzierung durch Heitzmann et al. (2019); sie bleiben von der Personalisierung unberührt.

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

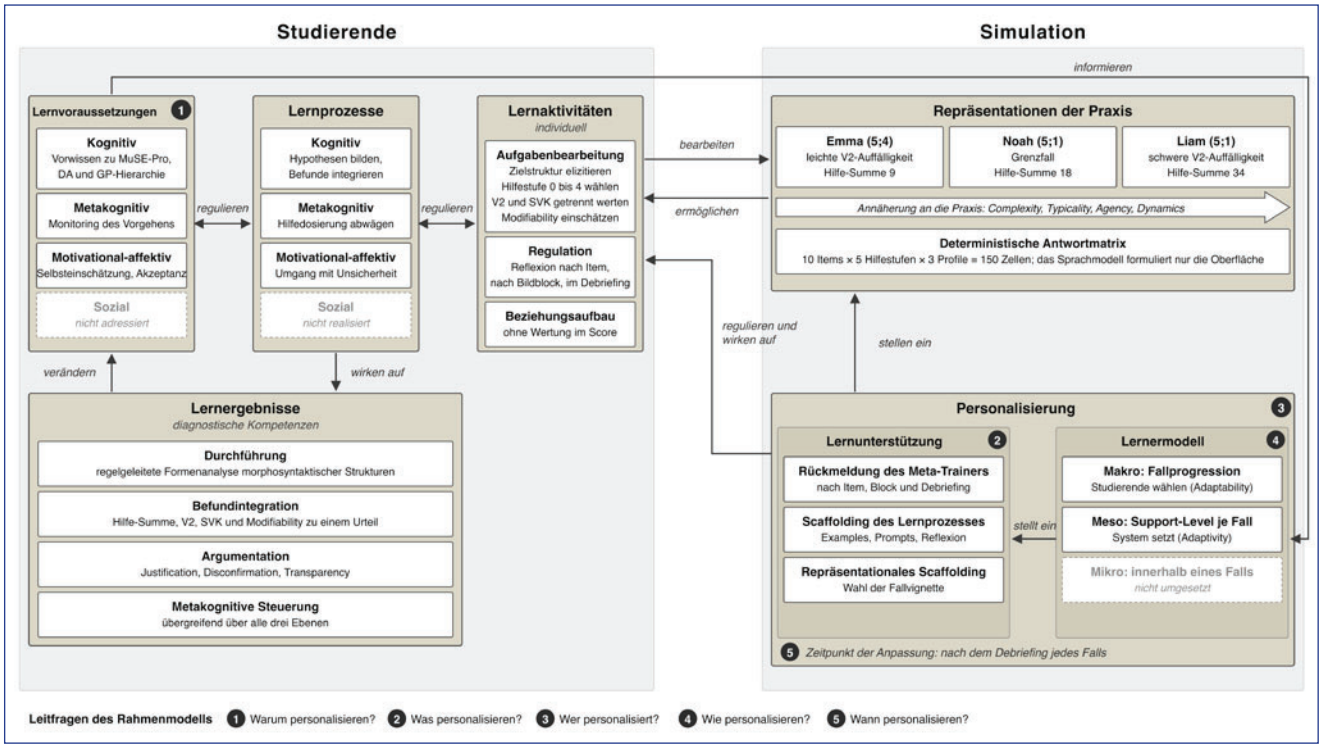


Abb. 1: Das SHARP-Rahmenmodell, bezogen auf die vorliegende Simulation.
Eigene Darstellung in Anlehnung an Bauer et al. (2025, Abb. 1): Die Struktur des Rahmenmodells ist übernommen, die Inhalte der Kästen und ihre Detaillierung stammen von den Autoren. Die Ziffern 1 bis 5 markieren, wo die fünf Leitfragen des Rahmenmodells beantwortet werden. Gestrichelt und grau gesetzt sind Elemente des Rahmenmodells, die die Simulation nicht umsetzt.

Zwei Gestaltungsachsen konkretisieren die Antworten auf die Fragen, was und wer personalisiert. Auf der Fallebene variieren die drei Fallvignetten entlang der vier Dimensionen des Representational Scaffolding nach Fischer et al. (2022), also in Komplexität, Typikalität, Handlungsspielraum und Dynamik. Auf der Ebene des Lernprozesses kombiniert der Meta-Trainer drei Scaffolding-Typen zu drei abgestuften Support-Leveln. Das höchste Level bietet Examples, Prompts und Reflexionsphasen gemeinsam an, das mittlere beschränkt sich auf Prompts und Reflexionsphasen und das niedrigste nutzt allein Reflexionsphasen. Maßgeblich für diese Abstufung ist die Metaanalyse von Chernikova et al. (2020b), weil sie sich auf diagnostische Kompetenzen bezieht. Sie weist Reflexionsphasen ($g = 0{,}58$ gegenüber 0,26 ohne; $p < .001$) und Prompts ($g = 0{,}47$ gegenüber 0,26; $p < .05$) als wirksame Moderatoren aus, Examples dagegen nicht ($g = 0{,}33$ gegenüber 0,39; n. s.). Dieselbe Metaanalyse berichtet jedoch eine Abhängigkeit vom Vorwissen: Reflexionsphasen wirken bei fortgeschrittenen Lernenden stärker, Beispiele bei weniger fortgeschrittenen. Genau dieser Befund begründet, warum das Lernermodell Examples nur bei geringem Vorwissen zuschaltet.
Die Personalisierung selbst folgt zwei orthogonalen Dimensionen. Adaptivity bezeichnet die systemseitige Anpassung auf Basis des Lernermodells, Adaptability die studierendenseitig gesteuerte Anpassung. Die Metaanalyse von Chernikova et al. (2025) zeigt eine differenzielle Wirksamkeit: Adaptivity wirkt vor allem beim instructional scaffolding, Adaptability vor allem bei der task progression. Die Simulation setzt beides getrennt um, indem das System das Trainer-Scaffolding adaptiv über das Lernermodell steuert, während die Studierenden die Fallprogression selbst wählen.

3 Konzeption der KI-gestützten Simulationsumgebung

Dieses Kapitel beantwortet den zweiten, gestaltungsbezogenen Teil der ersten Fragestellung. Es beschreibt, wie die Simulation aus dem in Kapitel 2 entwickelten Rahmen heraus gestaltet wurde, zuerst didaktisch (3.1 und 3.2), dann technisch (3.3 und 3.4). Die in Kapitel 4 berichtete Evaluation bezieht sich auf genau diesen Entwicklungsstand.

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

3.1 Didaktisches Ziel und Kompetenzfokus

Die Lernsimulation zielt auf den Aufbau diagnostischer Kompetenzen im Sinne der in Kapitel 2.1 entfalteten Begriffsbestimmung (Heitzmann et al., 2019). Die dort unterschiedenen drei Ebenen konkretisieren sich für die Diagnostik grammatischer Fähigkeiten in drei Teilkompetenzen: Durchführung, Befundintegration und Argumentation, ergänzt durch eine vierte, übergreifende metakognitive Steuerung. Die erste Teilkompetenz umfasst die regelgeleitete Formenanalyse morphosyntaktischer Strukturen, also das Erkennen von Verbzweitstellung (V2) und Subjekt-Verb-Kongruenz (SVK) einschließlich profiltypischer Produktionsmuster wie Verbendstellung oder Infinitivendstellung (Berg, 2015). Warum die Eingrenzung auf genau diese beiden Strukturen fällt, begründet Abschnitt 3.2. Die zweite ist die kriteriengeleitete Befundintegration, die Hilfe-Summenscore, V2- und SVK-Werte sowie Modifiability zu einem konsistenten diagnostischen Gesamtbild verdichtet (Beushausen & Walther, 2010). Die dritte ist die regelgeleitete Ableitung von Therapiezielen, in der sich die diagnostische Argumentation praktisch fortsetzt: Die begründete Befundlage wird in eine fachlich fundierte Indikation überführt (Spreer, 2018). Übergreifend wirkt die metakognitive Steuerung des eigenen Vorgehens, insbesondere das Monitoring der Hypothesenbildung und die Reflexion über die Hilfedosierung im Graduated Prompting (Ehlert, 2021a; Klein et al., 2019). Die KI fungiert in dieser Architektur als strukturierter Reflexionspartner, nicht als Antwortinstanz. Sie darf das diagnostische Reasoning weder vorwegnehmen noch ersetzen, da sonst gerade die Aktivitäten von der KI übernommen würden, die eigentlich trainiert werden sollen (Bauer et al., 2025). Stattdessen externalisiert sie das Reasoning schrittweise und initiiert Reflexion an drei strukturell relevanten Übergängen: nach einzelnen Items, beim Wechsel vom W-Fragen-Subtest zu den Inversionen und beim Eintritt in das Debriefing. Didaktisches Ziel ist damit ein Zuwachs an diagnostischer Kompetenz. Ob die Simulation ihn tatsächlich bewirkt, ist keine Frage des vorliegenden Beitrags: Die in Kapitel 4 berichtete Expertinnenevaluation prüft die fachlich-didaktische Tauglichkeit des Prototyps, nicht seine Lernwirksamkeit. Das dafür nötige Prä-Post-Design ist Gegenstand des nächsten DBR-Zyklus (vgl. 5.4).

3.2 Anwendungsszenario: Diagnostik grammatischer Fähigkeiten

Als prototypischer Anwendungsfall für die in der Forschungsfrage adressierten grammatischen Fähigkeiten dient die Diagnostik der Verbzweitstellung (V2) im Deutschen mittels DA. Dieser Gegenstand wurde in Abstimmung mit einer Expertinnenrunde gewählt, weil er linguistisch eindeutig und damit deterministisch operationalisierbar ist, eine Voraussetzung, auf der die in 3.3 beschriebene Architektur aufbaut. Als Grundlage für die Itemauswahl dient das sprachtherapeutische Testverfahren Morphologische und Syntaktische Entwicklung – Produktion (MuSE-Pro; Berg, 2015), aus dem zwei Untertests übernommen wurden, die die Verbzweitstellung mit je fünf Items elizitieren: einmal über W-Fragen, einmal über ein Temporaladverb in Erstposition. Das Kind soll dabei in beiden Fällen die Verbzweitstellung realisieren: Bei den W-Fragen steht zuerst das Fragewort, dann das finite Verb und anschließend das Subjekt, zum Beispiel: *„Wo wohnst du?“* Bei den Inversionen steht zuerst ein Temporaladverb, dann das finite Verb und anschließend das Subjekt, zum Beispiel: *„Jetzt schläft er.“*

Die Studierenden bearbeiten vier Anforderungen parallel. Sie evozieren die zehn Zielstrukturen und setzen bei nicht zielkonformen Antworten die vierstufige Hilfehierarchie des GP ein. Zugleich beobachten sie V2- und SVK-Korrektheit getrennt (Berg, 2015) und schätzen fortlaufend die Modifiability ein. Die Hilfehierarchie ist aufsteigend geordnet: Auf einen metakognitiven Hinweis zur Wortstellung *(„Wo muss das Verb im Satz stehen?“)* folgt eine Alternativfrage mit kontrastierenden Satzmustern *(„Heißt es ‚Jetzt der Junge schläft‘ oder ‚Jetzt schläft der Junge‘?“)*. Danach wird ein Satzanfang vorgegeben *(„Jetzt schläft …“)*. Als stärkste Unterstützung folgt ein vollständiges Modell mit Imitationsaufforderung *(„Jetzt schläft der Junge. Sag du es auch noch einmal.“)*. Diese Hilfehierarchie adaptiert das Stufenmodell von Ehlert (im Druck) und greift das aus Hasson et al. (2012) bekannte Prinzip auf, innerhalb einer Stufe mehrere gleichwertige Formulierungsvarianten bereitzuhalten. Ehlert (im Druck) bündelt Alternativfrage und Satzanfangsvorgabe gemeinsam auf der mittleren Stufe; hier werden sie zugunsten einer feineren diagnostischen Granularität getrennt. Gegenüber der Hierarchie der zugrunde liegenden Studie (Ehlert, 2021a; vgl. Tab. 6.4) entfällt zudem die Rezeptionsprüfung, die bei der V2-Inversion in dieser Altersgruppe kein tragfähiges Grammatikalitätsurteil erlaubt. Die drei Fallvignetten bilden unterschiedliche Grade von Responsivität ab: Emma zeigt eine hohe Modifiability und benötigt nur wenige Hilfen, Noah reagiert teilweise auf Unterstützung, Liam bleibt trotz stärkerer

Hilfen wenig veränderbar. Bei regelkonformer Durchführung wird dieser Unterschied über die kumulierten Hilfepunkte sichtbar. Dabei gilt: je mehr Unterstützung nötig ist, desto höher der Wert (vgl. Tab. 1).

Tab. 1: Die drei Fallvignetten der Simulation

Profil	Illustratives klinisches Bild	Erwartete Modifiability	Kalibrierte Hilfe-Summe
Emma	leichte V2-Auffälligkeit, günstige Prognose	Gut	9
Noah	deutliche V2-Auffälligkeit, Grenzfall	Mittel	18
Liam	schwere V2-Auffälligkeit, durchgängige Infinitivendstellung	Gering	34

Anmerkung: Die Modifiability-Erwartung ist systemseitig hinterlegt, wird den Studierenden aber nicht angezeigt; sie ist von ihnen aus dem Verhalten des Kindes zu erschließen. Die kalibrierte Hilfe-Summe ist der Wert, den eine regelkonforme Durchführung ohne übersprungene Stufen ergibt (Maximum 40).

Worin sich die Profile konkret unterscheiden, zeigt ein Item der Antwortmatrix (vgl. Tab. 2). Für jede Kombination aus Item, Hilfestufe und Profil ist dort festgelegt, welche Äußerung das Kind produziert und wie sie gewertet wird. Emma und Noah stellen das finite Verb ans Satzende und sind damit in der Verbzweitstellung auffällig, nicht aber in der Subjekt-Verb-Kongruenz; Liam verwendet durchgängig einen Infinitiv in Endstellung und verfehlt beide Zielmerkmale. Emma und Noah unterscheiden sich nicht in diesem Item, sondern darin, auf wie vielen der zehn Items sie die Zielstruktur ohne Hilfe produzieren; daraus ergeben sich die Hilfe-Summen 9 und 18. Die Zellen wurden aus dem Testmanual und der DA-Literatur abgeleitet (Berg, 2015; Ehlert, 2021a; Hasson et al., 2012); ihre Plausibilität war Gegenstand der in Kapitel 4 berichteten Evaluation.

Tab. 2: Ausschnitt aus der Antwortmatrix für Item 6 (Zielstruktur „Zuerst schläft der Hund.“)

Angebotene Hilfestufe	Emma	Noah	Liam
0 keine Hilfe	*Zuerst der Hund schläft.* (V2 –, SVK +)	*Zuerst der Hund schläft.* (V2 –, SVK +)	*Zuerst der Hund schlafen.* (V2 –, SVK –)
1 metakognitiver Hinweis	*Zuerst der Hund schläft?* (V2 –, SVK +)	*Zuerst der Hund schläft.* (V2 –, SVK +)	*Zuerst der Hund schlafen.* (V2 –, SVK –)
2 Alternativfrage	*Ähm, zuerst der Hund schläft.* (V2 –, SVK +)	*Zuerst der Hund schläft.* (V2 –, SVK +)	*Zuerst der Hund schlafen.* (V2 –, SVK –)
3 Satzanfang vorgeben	***Zuerst schläft der Hund!*** ***(V2 +, SVK +)***	***Zuerst schläft der Hund.*** ***(V2 +, SVK +)***	*Zuerst der Hund schlafen.* (V2 –, SVK –)
4 vollständiges Modell	nicht erreicht	nicht erreicht	***Zuerst schlafen der Hund.*** ***(V2 +, SVK –)***

Anmerkung: V2 = Verbzweitstellung, SVK = Subjekt-Verb-Kongruenz; + korrekt, – inkorrekt. Ein infiniter Verbstamm wie „schlafen“ gilt als SVK-inkorrekt und in Zweitposition zugleich als V2-korrekt.
Fett markiert ist die jeweilige Lösungsstufe, auf der das Item schließt und deren Wert in die Hilfe-Summe eingeht (3, 3 und 4).

3.3 Technische Umsetzung: deterministische Antwortmatrix und Sprachmodell

Die bisher beschriebenen Designentscheidungen stellen eine Anforderung an die technische Umsetzung, die didaktisch begründet ist. Eine Fallvignette ist in dieser Simulation kein Erzählmaterial, sondern ein Messinstrument: Die Hilfe-Summe, die eine Studierende am Ende notiert, ist nur dann interpretierbar und mit der einer Kommilitonin vergleichbar, wenn das Kind in derselben Situation dieselbe Struktur produziert hat. Daraus folgt die Frage, die dieser Abschnitt beantwortet: Wie stellt eine sprachmodellbasierte Umgebung sicher, dass alle Studierenden denselben Fall bearbeiten?

Dass diese Frage nicht trivial ist, zeigten eigene Vorarbeiten. In Erprobungen eines ersten Prototyps, in dem ein einzelner KI Agent Kindverhalten und Trainerrückmeldung zugleich erzeugte, wich das Sprachmodell wiederholt vom hinterlegten Profil ab: Es gab profilfremde Antworten und korrigierte diagnostisch tragende Produktionsmuster unaufgefordert, sodass das Kind gera-

de jene Auffälligkeit nicht mehr zeigte, die die Studierende erkennen sollte. Dieses Verhalten ist kein Einzelfall des hier verwendeten Modells. Li et al. (2024) zeigen in einem Benchmark aus simulierten Dialogen, dass Sprachmodelle bereits über einen Horizont von acht Gesprächsrunden von der zu Beginn gegebenen Instruktion abweichen; geprüft wurde das an LLaMA2-chat-70B und GPT-3.5. Die Architektur lagert die diagnostisch tragende Strukturentscheidung deshalb in eine deterministische Antwortmatrix aus und reduziert das Sprachmodell auf die sprachliche Oberfläche. Sie folgt damit der Forderung von Bauer et al. (2025), dass KI-Werkzeuge zentrale Lernaktivitäten ermöglichen und nicht ersetzen sollen.

Die Simulation ist als Agentflow auf der Plattform Flowise realisiert (vgl. Abb. 2), also als visuell modellierter Arbeitsablauf, in dem jeder Verarbeitungsschritt als eigener Baustein vorliegt und alle Bausteine auf einen gemeinsamen, über die Sitzung fortgeschriebenen Datensatz zugreifen. Zu Beginn einer Sitzung übernimmt der Ablauf die gewählte Fallvignette und das aus dem Vortest bestimmte Support-Level. Anschließend entscheidet bei jedem Beitrag der Studierenden eine Klassifikationskomponente, welcher von drei Pfaden ihn verarbeitet. Welcher Pfad das ist, lässt sich an drei prototypischen Beiträgen veranschaulichen. Eine Itemelizitation wie *„Frag das Tier doch mal unter der Decke, wo es wohnt“* leitet auf den Diagnostikpfad, ein Beziehungsbeitrag wie *„Hey Emma, wie geht es dir heute?“* auf den Beziehungsaufbau-Pfad und eine Verfahrensfrage wie *„Welche Hilfestufe wäre hier sinnvoll und warum?“* auf den Meta- und Debriefingpfad. Eine solche Gliederung der Begegnung in unterscheidbare Teile, etwa Vorbesprechung, Szenario und Debriefing, ist in Simulationsframeworks üblich (Eppich & Cheng, 2015; Hewat et al., 2020). Die vorliegende Architektur überträgt dieses Prinzip auf die Ebene des einzelnen Beitrags. Beiträge des Beziehungsaufbaus bleiben von der Score-Logik ausgenommen, während Verfahrensfragen und die abschließende Diagnostik auf den Meta- und Debriefingpfad gehen.

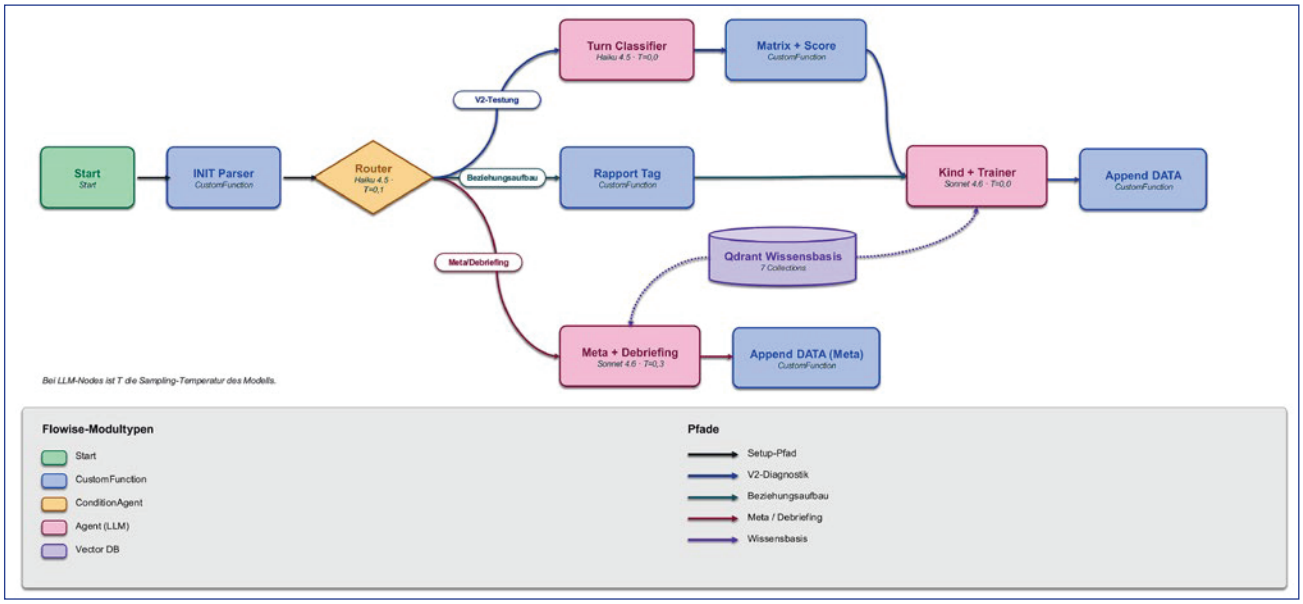


Abb. 2: Flowise-Agentflow der KI-gestützten DA-Trainingssimulation. Modulart pro Knoten farblich markiert; die drei Pfade (V2-Diagnostik, Beziehungsaufbau, Meta/Debriefing) sind über die Pfeilfarben referenziert; die Wissensbasis ist über gestrichelte Verbindungen an die zwei Agent-Nodes Kind + Trainer und Meta + Debriefing angebunden.

Auf dem V2-Diagnostikpfad entscheidet nicht das Sprachmodell, sondern eine feste, regelbasierte Programmroutine, welche Antwort das Kind gibt. Diese bewusste Trennung verhindert, dass das Sprachmodell diagnostisch tragende Produktionsmuster eigenmächtig erzeugt oder verändert, und stellt sicher, dass die grammatische Struktur der Kindantwort in jedem Studierendendurchlauf identisch bleibt. Der Baustein greift auf die hinterlegte Antwortmatrix zu, die pro Kombination aus Item, GP-Hilfestufe und Profil eine Zelle vorhält. Jede Zelle enthält den Strukturcode der erwarteten Kindantwort, deren V2- und SVK-Korrektheit für Elizitationen ohne vorgegebenes Zielverb sowie das profiltypische Modifiability-Verhalten. Tabelle 2 zeigt diese Zellen für ein Item. Erst danach erzeugt der Kind + Trainer-Agent die sprachliche Oberfläche, gebunden an den festgelegten Strukturcode. Das Sprachmodell variiert ausschließlich die Wortwahl im profil-

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

typischen Korridor und passt diese an die vorausgegangene Elizitationsstruktur der Studierenden an, die zuvor durch den Turn-Classifier (Äußerungsklassifikator) extrahiert wurde. In Pilottests blieb die Strukturtreue über mehrere hundert Turns matrixkonform. Eine Wissensbasis bindet die Reflexion des Trainers an hinterlegte Fachliteratur an, sodass sie nicht ausschließlich aus dem parametrischen Wissen des Sprachmodells gespeist wird.

3.4 Benutzeroberfläche und Sitzungsablauf

In einer gewöhnlichen Chat-Oberfläche würden Kindäußerung und Trainer-Kommentar als ein einziger Textfluss erscheinen. Studierende müssten dann vor jedem diagnostischen Schritt entscheiden, welcher Satz das Kind zeigt und welcher ein didaktischer Impuls ist. Diese Vorab-Klassifikation erzeugt aufgabenfremde kognitive Last (Klepsch et al., 2017) zusätzlich zu einer ohnehin anspruchsvollen Aufgabe. Die Anwendungsschicht der Simulation verdichtet die Ausgabeblöcke der in 3.3 beschriebenen Backend-Architektur zu einer lernwirksamen Sitzung. Den unmittelbar wirksamen Hebel bildet die visuell-auditive Trennung der Sprecherrollen [KIND] und [TRAINER]. In einer eigens entwickelten HTML-Anwendung erscheinen beide als getrennte, pro Sprecherrolle farblich unterschiedene Sprechblasen (vgl. Abb. 3). Die Kindzeile wird zusätzlich über die Voice-Synthesis-Engine ElevenLabs (ElevenLabs Inc., 2024) auditiv ausgegeben, sodass profiltypische Sprachmerkmale wie Tempo, Lautheit und Pausenverhalten in derselben Modalität wie im realen Kontakt erfahrbar werden. Spiegelbildlich transkribiert OpenAI Whisper (Speech-to-Text; OpenAI, 2022) die Studierenden-Elizitationen.

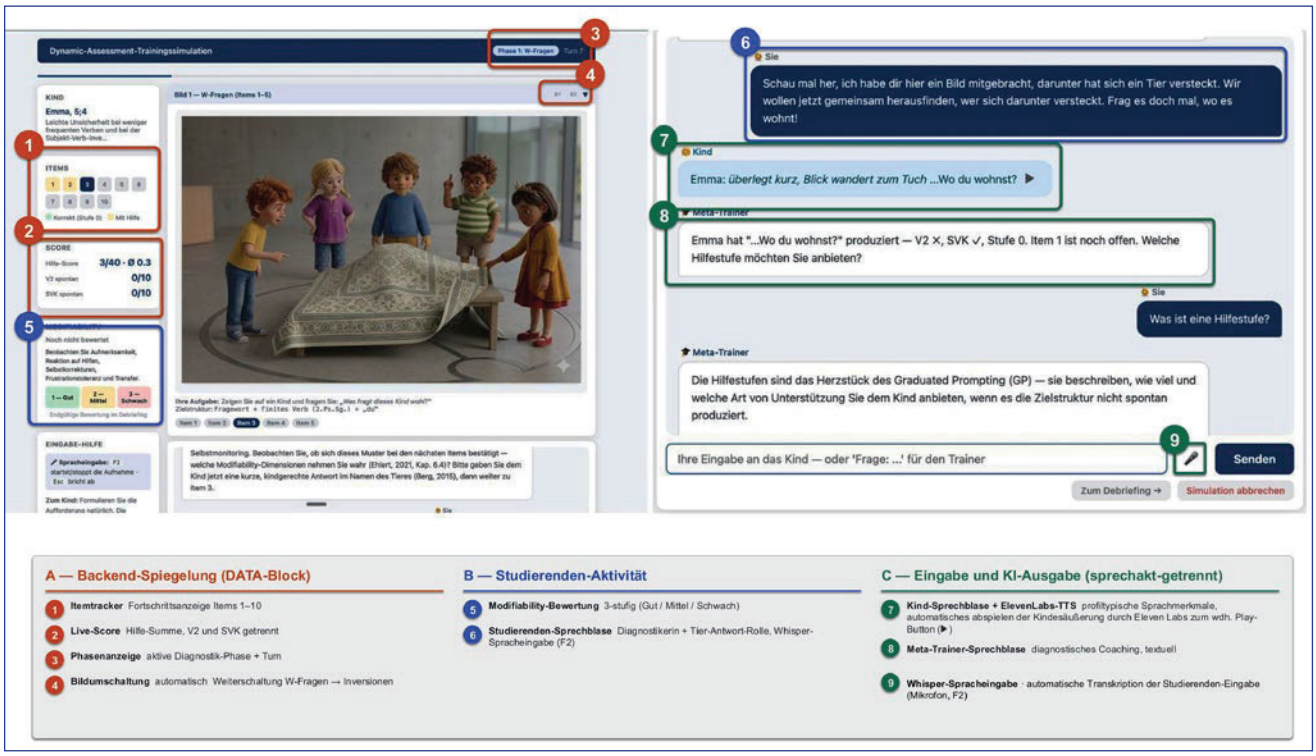


Abb. 3: Hauptdialog des HTML-Frontends.

Die Oberfläche zeigt den Stand der Testung fortlaufend an. Dazu dienen vier Anzeigen: ein Itemtracker über die zehn Items, ein fortlaufender Hilfe-Summenscore mit getrennt ausgewiesenen V2- und SVK-Werten, der automatische Bildwechsel vom W-Fragen-Subtest zu den temporalen Inversionen sowie der Übergang in die Debriefingansicht, in der eine vom Meta-Trainer geleitete und an das Support-Level angepasste Diagnosereflexion stattfindet.

In einer Sitzung verbindet die Simulation Vortest, theoretische Einführung, Fallbriefing, Hauptdialog, Debriefing, und Posttest in einen durchgehenden Ablauf ohne Wechsel der Simulation. Der Vortest erhebt das Lernermodell, auf dessen Grundlage das Sprachmodell die Sitzung adaptiv ausgestaltet. Nach dem Debriefing eines Falls wählen die Studierenden selbst, ob sie den Fall mit verändertem Support-Level wiederholen, zur nächsten Vignette wechseln oder einzelne Items re-

flektieren. Der Posttest schließt mit der satisfaction with simulation experience scale (SSES) nach Levett-Jones et al. (2011) als Akzeptanzindikator und einem Cognitive-Load-Inventar (CLT-Q; Klepsch et al., 2017) als Belastungsindikator.

4 Explorative Expertinnenstudie zur Umsetzbarkeit und zum Mehrwert der KI-Simulationsumgebung

4.1 Zielsetzung, Fragestellung und Studiendesign

Die explorative Expertinnenstudie zielte auf eine erste formative Einschätzung der entwickelten KI-gestützten Simulationsumgebung für die sprachtherapeutische Hochschullehre. Zur Beantwortung der in Kapitel 1 formulierten Evaluationsfrage (Forschungsfrage 2) wurde folgende Teilfragestellung konkretisiert: Wie beurteilen sprachtherapeutische Expertinnen die Umsetzbarkeit und den didaktischen Mehrwert der Simulationsumgebung für den Einsatz im Studiengang Sprachtherapie?
Methodisch ist die Studie im Design-Based-Research-Ansatz (DBR) verortet. Dieser eignet sich für Entwicklungsprojekte, die theoretisch fundierte Innovationen für konkrete Bildungskontexte hervorbringen und iterativ verbessern wollen (Barab, 2006; Ford et al., 2017). Reinmann (2005) charakterisiert DBR als Ansatz, der Gestaltung, Erprobung, Analyse und Re-Design zyklisch verbindet und auf nachhaltige Innovationen im Bildungsalltag zielt. Diesem Ansatz folgend wurde der Prototyp, der auf den theoretischen Vorarbeiten zu DA, diagnostischer Kompetenz, fallbasiertem Lernen und KI-gestützten Simulationen aufbaut (Kapitel 2–3), in einer Expertinnenevaluation (n = 3) fachlich-didaktisch eingeschätzt. Die Evaluation kombinierte eine Fragebogenerhebung vor und nach der Arbeit mit der Lernumgebung mit einem anschließenden halbstrukturierten Interview. So ließen sich sowohl der Erwartungshorizont vor der Nutzung als auch erfahrungsbezogene Rückmeldungen nach der konkreten Erprobung des Prototyps erfassen. Die Ergebnisse dienen der iterativen Weiterentwicklung der Simulationsumgebung mit dem Ziel, ein praktikables und fachlich fundiertes Fördertool für die Hochschullehre bereitzustellen.

4.2 Stichprobe, Erhebungssituation und Erhebungsinstrumente

Die Erhebung erfolgte in zwei Online-Sitzungen über Zoom, jeweils mit Einverständniserklärung zur Aufzeichnung. Sitzung 1 (90 Minuten) führte die Expertinnen A und B parallel im selben Konferenzraum zusammen, jede in einer eigenen Simulationssitzung; Sitzung 2 mit Expertin C fand wenige Tage später im Einzelsetting statt (60 Minuten). Die drei Expertinnen sind aktuelle oder ehemalige wissenschaftliche Mitarbeiterinnen des Lehrstuhls für Sprachheilpädagogik (LMU München). Da die Einarbeitung in Sitzung 1 als komplex erlebt worden war, wurde das Skript für Sitzung 2 um eine dreiminütige Live-Demonstration des Werkzeug-Interfaces ergänzt, eine bewusste Iteration im Sinne der Zyklenlogik des DBR (Reinmann, 2005).

Die Auswahl der Expertinnen beruhte auf drei Kriterien. Maßgeblich waren Lehrerfahrung in der sprachdiagnostischen Ausbildung des Bachelorstudiengangs, Erfahrung mit der Diagnostik grammatischer Fähigkeiten einschließlich MuSE-Pro (Berg, 2015) sowie Vertrautheit mit dem Dynamic Assessment. Die drei Expertinnen sind promovierte Sprachtherapeutinnen mit neun bis vierzehn Jahren Erfahrung in der sprachtherapeutischen Hochschullehre. Sie lehren oder lehrten Diagnostikmodule im Studiengang Sprachtherapie an der Ludwig-Maximilians-Universität München, zwei von ihnen zum Erhebungszeitpunkt als wissenschaftliche Mitarbeiterinnen des Lehrstuhls, die dritte als ehemalige Mitarbeiterin. Neben der Hochschultätigkeit sind oder waren sie klinisch tätig, in eigener Praxis, in der Frühförderung oder in der Beratungsstelle des Lehrstuhls. Die institutionelle Nähe erleichterte den Zugang zu einschlägiger Expertise, begrenzt aber zugleich die Unabhängigkeit der Urteile; beides wird in Abschnitt 5.4 aufgegriffen.

Die Erhebung erfolgte nach schriftlicher Aufklärung und Einwilligung. Die Expertinnen stimmten der Aufzeichnung, der Transkription und der pseudonymisierten Verwendung ihrer Aussagen gesondert zu; die Zuordnung erfolgt im Beitrag ausschließlich über die Kennungen A, B und C. Die weiterreichenden Fragen, die der Einsatz generativer KI mit Studierenden aufwirft, erörtert Abschnitt 5.3. Tabelle 3 stellt eine Auflistung der eingesetzten Erhebungsinstrumente mit Konstrukt, Itemzahl und Skala dar, geordnet nach den vier Erhebungsschritten.

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

Tab. 3: Eingesetzte Erhebungsinstrumente

	Instrument	Konstrukt / Subskala	Items	Skala
1	**Vor der Erprobung, Bewertungsbogen Teil 1**			
	UTAUT2 (Venkatesh et al., 2012)	Technologieakzeptanz	4	1–5
	Erprobung: Fallvignette „Emma“ im Modus „High Support“, 16,3–23,3 Minuten, keine Erhebung			
2	**Direkt nach der Erprobung, in der Simulation**			
	CLT-Q (Klepsch et al., 2017)	Kognitive Belastung (Intrinsic, Extraneous, Germane)	9	0–8
3	**Nach der Erprobung, Bewertungsbogen Teil 2**			
	SSES (Levett-Jones et al., 2011)	Clinical Learning, Clinical Reasoning, Debriefing & Reflection	9	1–5
	SUS-Adaption (Brooke, 1996)	System-Usability	3	1–5
	Projektspezifische Items	GP-Hilfehierarchie, Modifiability-Beobachtung, integratives Urteil, Meta-Trainer-Präferenz, Curriculum-Integration	5	1–5
	Zielgruppen und Einsatzkontexte	Nutzungsverortung	Mehrfachnennung	nominal
	Offene Rückmeldungsfelder	qualitatives Feedback	4	offen
4	**Abschluss**			
	Reflexionsgespräch	leitfadengestützt	8 Leitfragen	offen

Die Erhebung folgte in beiden Sitzungen demselben Ablauf aus vier Erhebungsschritten, die sich um die Erprobung der Simulation gruppieren (vgl. Tab. 3). Den ersten Schritt bildete noch vor der Erprobung der erste Teil eines schriftlichen Bewertungsbogens mit vier UTAUT2-Items zur Technologieakzeptanz (Venkatesh et al., 2012). Darauf folgte die Erprobung der Fallvignette „Emma“ im Modus „High Support“ (durchschnittlich 19,6 Minuten, Spanne 16,3–23,3), unterstützt durch ein Begleit-Manual mit Interface-Instruktionen und Hilfehierarchie. Sie ist der Gegenstand der Evaluation und selbst kein Erhebungsschritt, weshalb während der Erprobung nicht erhoben wurde. Den zweiten Schritt bildete unmittelbar danach, noch in der Simulation, die Erfassung der kognitiven Belastung mit neun Items in den Subskalen Intrinsic, Extraneous und Germane (CLT-Q; Klepsch et al., 2017). Den dritten Schritt bildete der zweite Teil desselben Bewertungsbogens, den alle drei Expertinnen nach der Erprobung ausfüllten. Er umfasste neun SSES-Items (Levett-Jones et al., 2011) zu den Subskalen Clinical Learning, Clinical Reasoning und Debriefing & Reflection, drei Items zur Gebrauchstauglichkeit (Brooke, 1996), fünf projektspezifische Items, Mehrfachnennungen zu Zielgruppen und Einsatzkontexten sowie vier offene Rückmeldungsfelder. Die drei Usability-Items bilden eine Kurzfassung der System Usability Scale (SUS; Brooke, 1996). Gewählt wurden diejenigen drei Items, die sich auf die hier interessierende Bedienbarkeit durch Studierende beziehen: die Vorstellbarkeit regelmäßiger Nutzung, die wahrgenommene Komplexität (invertiert gewertet) und die Einfachheit der Bedienung. Die übrigen Items der Skala zielen auf Systemeigenschaften, die in einem Prototyp ohne Produktivbetrieb nicht beurteilbar sind. Da diese Kurzfassung nicht eigens validiert wurde, werden die drei Items deskriptiv und ausdrücklich nicht als SUS-Score berichtet.

Den vierten Schritt und Abschluss bildete ein leitfadengestütztes Reflexionsgespräch von etwa 20 bis 30 Minuten, in Sitzung 1 gemeinsam mit beiden Expertinnen, in Sitzung 2 im Einzelsetting. Acht Leitfragen strukturierten das Reflexionsgespräch in flexibler Reihenfolge:
1. Ersteindruck vom simulierten Kind,
2. erlebte Rolle des Meta-Trainers,
3. klinischer Realismus,
4. didaktischer Transfer in die eigene Lehre einschließlich Zielgruppe und Einsatzkontext,
5. Mehrwert gegenüber klassischen Fallformaten,
6. Sinnhaftigkeit der Meta-Trainer-Begleitung gegenüber einem reinen Kind-Dialog,
7. größte Hürde für den Einsatz,
8. offener Abschluss für nicht angesprochene Punkte.

Die Leitfragen greifen teilweise dieselben Konstrukte auf wie der Bewertungsbogen. So werden Zielgruppen und Einsatzkontexte doppelt erfasst, einmal geschlossen über die Mehrfachnennung

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

im Bewertungsbogen und einmal offen über Leitfrage 4. Diese bewusste Überlappung ermöglicht die in Abschnitt 4.3 beschriebene konstruktweise Zusammenführung von Fragebogen- und Interviewdaten.

4.3 Datenauswertung

Die qualitative Auswertung der Audiotranskripte folgt der strukturierenden Inhaltsanalyse nach Mayring (2022). Das Kategoriensystem wurde deduktiv aus den vier Bewertungskonstrukten abgeleitet, die auch den Bewertungsbogen strukturieren: inhaltliche Validität, theoretische Treue, klinischer Realismus und didaktischer Wert. Die Konstrukte lehnen sich an Rahmenmodelle zur Evaluation simulationsbasierter Lernumgebungen an (Bauer et al., 2025; Chernikova et al., 2020a). Hinzu kamen drei Querthemen, die sich aus dem Erhebungssetting ergaben, nämlich technische Beobachtungen, Bedienbarkeit und Voraussetzungen auf Seiten der Nutzenden, sowie eine induktiv gebildete Restkategorie für nicht zuordenbare Aussagen.

Zu jeder Kategorie hält ein Kodierleitfaden eine Definition, mindestens ein Ankerbeispiel aus dem Material und eine Kodierregel bereit, die die Abgrenzung zu benachbarten Kategorien festlegt (Mayring, 2022). Als Selektionskriterium galt, dass eine Aussage einen erkennbaren Bezug zum Werkzeug oder zu seinem Lehreinsatz haben muss; reine Verständigungs- und Organisationsbeiträge, etwa zu Zoom-Verbindung oder Terminabsprache, blieben unkodiert. Kodiereinheit war die inhaltlich abgeschlossene Aussage, Kontexteinheit der jeweilige Gesprächsabschnitt. Nach einer Probekodierung wurden Kategoriensystem und Kodierregeln überarbeitet. Insgesamt wurden 88 Aussagen kodiert. Die Kodierung erfolgte durch den Erstautor; auf eine quantifizierte Intercoderreliabilität wurde angesichts der Stichprobengröße verzichtet (vgl. 5.4). Zur Gütesicherung wurden Kategoriensystem und strittige Einzelentscheidungen diskursiv abgestimmt (kommunikative Validierung, Mayring, 2022).

Die Likert-Items wurden deskriptiv ausgewertet. Weil die Leitfragen des Reflexionsgesprächs dieselben Konstrukte adressieren wie der Bewertungsbogen, ließen sich beide Datenarten konstruktweise nebeneinanderlegen: Für jedes Konstrukt wurde geprüft, ob die kodierten Aussagen in dieselbe Richtung weisen wie die zugehörigen Likert-Werte. Übereinstimmungen werden im Folgenden als konvergente Befunde berichtet, Abweichungen zwischen beiden Datenarten eigens ausgewiesen. Eine statistische Verrechnung der beiden Datenarten fand nicht statt.

4.4 Ergebnisse der Expertinnenstudie

Der Bewertungsbogen wurde auf einer fünfstufigen Likert-Skala erhoben (1 bis 5, höhere Werte günstiger), die CLT-Q-Subskalen auf einer neunstufigen Skala (0 bis 8, höhere Werte mehr Belastung). Angesichts von n = 3 berichtet Tabelle 4 die Werte der drei Expertinnen einzeln sowie den Median. Auf Mittelwerte über drei Personen wird verzichtet, weil sie eine Messgenauigkeit suggerieren, die die Erhebung nicht hergibt. Aus den 88 kodierten Aussagen und den Likert-Werten ergeben sich vier Befunde: drei betreffen das Werkzeug, einer die Voraussetzungen auf Seiten der Nutzenden.

Der voraussetzungsseitige Befund betrifft eine Vorbereitungs-Hürde. Alle drei Expertinnen knüpfen die Wirksamkeit des Werkzeugs unabhängig voneinander an drei Komponenten: Vertrautheit mit MuSE-Pro (Berg, 2015), DA und GP sowie Geübtheit im Umgang mit der Bedienoberfläche. Am deutlichsten formuliert es Expertin C im offenen Feld: Es sei zwingend notwendig, MuSE-Pro und Dynamic Assessment mit den Studierenden vorab ausführlich zu besprechen; die Simulation sei dann hilfreich, wenn sich die Studierenden mit beidem schon gut auskennen und es um die Erprobung der Anwendung gehe. Die CLT-Q-Werte sind damit vereinbar, erlauben aber keine kausale Deutung. In Sitzung 1 lag die intrinsische Belastung bei 3,67 und 7,00, in Sitzung 2 nach der ergänzten Live-Demonstration bei 1,33. Der Abstand zwischen den beiden Expertinnen der ersten Sitzung ist dabei fast so groß wie der Abstand zur zweiten Sitzung, sodass interindividuelle Unterschiede als Erklärung nicht auszuschließen sind. Aussagekräftiger erscheint das Gesamtprofil der zweiten Sitzung: niedrige intrinsische und niedrige aufgabenfremde Belastung bei zugleich der höchsten lernbezogenen Belastung aller drei Erhebungen (6,33). Dieses Profil entspricht einer von Klepsch et al. (2017) als lernförderlich beschriebenen Konstellation, lässt sich bei einer einzelnen Erhebung aber nicht belastbar interpretieren.

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

Tab. 4: Zentrale Bewertungskennwerte der Expertinnenevaluation

Konstrukt (Instrument)		A	B	C	Median
1	**Vor der Erprobung, Bewertungsbogen Teil 1, Skala 1–5 (höhere Werte günstiger)**				
	Technologieakzeptanz (UTAUT2, 4 Items)	4,00	3,75	3,75	3,75
Erprobung: Fallvignette „Emma“ im Modus „High Support“, 16,3–23,3 Minuten, keine Erhebung					
2	**Direkt nach der Erprobung, in der Simulation, Skala 0–8 (höhere Werte mehr Belastung)**				
	Intrinsische Belastung (CLT-Q)	3,67	7,00	1,33	3,67
	Aufgabenfremde Belastung (CLT-Q)	3,33	4,00	1,33	3,33
	Lernbezogene Belastung (CLT-Q)	4,00	3,00	6,33	4,00
3	**Nach der Erprobung, Bewertungsbogen Teil 2, Skala 1–5 (höhere Werte günstiger)**				
	Clinical Learning (SSES, 3 Items)	4,33	4,33	4,00	4,33
	Clinical Reasoning (SSES, 3 Items)	4,67	4,67	5,00	4,67
	Debriefing & Reflection (SSES, 3 Items)	5,00	5,00	3,67	5,00
	System-Usability (SUS-Adaption, 3 Items)	3,00	3,67	3,67	3,67
	– darunter Einzelitem „einfach zu bedienen“	2	2	4	2
	P1 Anwendung der GP-Hilfehierarchie	4	5	3	4
	P2 Modifiability-Beobachtung	3	5	3	3
	P3 Integratives klinisches Urteil	3	5	3	3
	P4 Meta-Trainer-Variante gegenüber reinem Kind-Dialog vorgezogen	2	1	2	2
	P5 Integrierbarkeit in ein bestehendes Curriculum	5	5	5	5

Anmerkung: A und B bearbeiteten die Simulation in Sitzung 1 parallel, C in Sitzung 2 nach der ergänzten Live-Demonstration. Werte mit zwei Nachkommastellen sind Subskalenmittel über die Items einer Person, einstellige Werte sind Einzelitems. Das Item zur wahrgenommenen Komplexität geht invertiert in die System-Usability ein. P1 bis P5 sind projektspezifische Einzelitems ohne psychometrische Validierung und werden deshalb nicht zu einer Skala verrechnet.

Die drei simulationsbezogenen Befunde betreffen didaktische Tauglichkeit, Bedienoberfläche und klinischen Realismus. Erstens fällt die didaktische Tauglichkeit unter erfüllten Vorbedingungen hoch aus: Clinical Reasoning liegt bei allen drei Expertinnen im oberen Skalenbereich, die Integrierbarkeit in ein bestehendes Curriculum bewerten alle drei mit dem Höchstwert (vgl. Tab. 4). Als Einsatzkontexte bestätigen alle drei die Prüfungsvorbereitung und die Vorbereitung auf reale Fallarbeit, die asynchrone Selbstlernphase zwei von drei. Bei den Zielgruppen gehen die Einschätzungen auseinander: Zwei Expertinnen verorten die Simulation in der Bachelor-Vertiefung, im Masterstudium und in der Fortbildung, die dritte ausschließlich in den Bachelor-Grundlagen, sodass sich keine einheitliche curriculare Verortung ergibt.

Zweitens markiert die Bedienoberfläche den deutlichsten Verbesserungsbereich (vgl. Tab. 4). Zwei Punkte kehren in den offenen Rückmeldungen übereinstimmend wieder: das Verschwinden des MuSE-Pro-Testbildes mit dem Chatverlauf und die Verteilung notwendiger Information auf Manual, Patientenakte und Oberfläche. Expertin A schlägt vor, Instruktionen und Hilfehierarchie in die Oberfläche zu integrieren und den Kinddialog sowie das Trainergespräch in getrennten Fenstern zu führen, statt beides auf einem separaten Blatt beziehungsweise in einem Fenster bereitzuhalten.

Drittens wird der klinische Realismus für die Kindfigur und für das V2-Produktionsmuster der Verbendstellung bestätigt. Uneinheitlich fällt das Urteil über die Trainerbegleitung aus; erst beide Datenarten zusammen lösen es auf. Die Subskala Debriefing & Reflection liegt hoch und weist die Trainerfunktion in den offenen Feldern als ausdrückliche Stärke aus. Expertin A nennt das individuelle Feedback eine „1:1-Betreuung“. Das direkt darauf gerichtete Item P4, ob die Variante mit Meta-Trainer einer reinen Kind-Dialog-Variante vorgezogen würde, beantworten jedoch alle drei ablehnend (2, 1 und 2). Das Reflexionsgespräch klärt den scheinbaren Widerspruch: Nicht die Rückmeldung selbst, sondern dass sie nach jedem Item erfolgt, wird als störend erlebt, weil sie den Gesprächsfluss unterbricht. Eine Expertin formulierte, man komme so „nicht so wirklich

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

in den Flow und in die natürliche Interaktion", eine zweite bezeichnete sich als unentschieden und verwies auf die zusätzliche Lesemenge. Aus der Gruppe heraus entstand der Vorschlag einer Mischform: ein erster Durchlauf ohne Unterbrechung, differenzierte Rückmeldungen erst in späteren Durchläufen.

5 Diskussion

5.1 Interpretation und didaktische Verortung der Expertinnenbefunde

Die Expertinnen urteilen differenziert statt pauschal. Die konzeptbezogenen Subskalen erreichen hohe Werte, Clinical Reasoning mit 4,67 bis 5,00 und die Curriculum-Integration bei allen dreien mit dem Höchstwert. Die Bewertung der Bedienoberfläche fällt demgegenüber deutlich ab (3,00 bis 3,67; Einzelitem zur Bedienbarkeit 2, 2 und 4). Didaktische Tragfähigkeit des Konzepts und Reife der Umsetzung werden also getrennt beurteilt. Für die Weiterentwicklung ist das die nützlichste Form der Rückmeldung, weil sie den Überarbeitungsbedarf lokalisiert, statt ihn über das Gesamturteil zu mitteln. Ein Beleg für die Validität des Instruments ist dieses Muster nicht; dafür wären eine größere und unabhängigere Stichprobe sowie ein Vergleichsmaßstab nötig.

Ein zweiter Befund illustriert die zyklische Logik des DBR (Reinmann, 2005). Nach der für Sitzung 2 ergänzten Live-Demonstration fiel die wahrgenommene Aufgabenschwierigkeit niedriger aus als in Sitzung 1, und das Belastungsprofil verschob sich zugunsten der lernbezogenen Komponente. Bei drei Personen und erheblicher Streuung innerhalb der ersten Sitzung ist das ein Hinweis, kein Wirknachweis; die Iteration ist damit begründet, ihre Wirkung aber nicht belegt. Die Studie realisiert einen ersten, in sich geschlossenen DBR-Zyklus. Die für einen Lernwirksamkeitsnachweis nötigen weiteren Zyklen, insbesondere die Pilotierung mit Studierenden, schließen daran an (vgl. 5.4).

Auf der Designebene beantwortet der Meta-Trainer die SHARP-Leitfrage, was personalisiert wird (Bauer et al., 2025): Er übersetzt eine personalintensive Diagnostikbegleitung in eine skalierbare Architektur. Die Expertinnenbefunde stützen diese Designorientierung in ihrem Kern und widersprechen ihr in der Ausführung. Gestützt wird die Funktion: Die Subskala Debriefing & Reflection liegt hoch. Qualitativ verdichtet sich die Rückmeldung in der von Expertin A geprägten Formel der „1:1-Betreuung", die sich in der Inhaltsanalyse nach Mayring (2022) als zentraler Kodierungsanker erwies. Widersprochen wird dem Eingreifen nach jedem Item: Dass alle drei Expertinnen die Meta-Trainer-Variante nicht einer reinen Kind-Dialog-Variante vorziehen, ist der deutlichste Einzelbefund der Studie. Er betrifft nicht die Frage, ob der Trainer sinnvoll ist, sondern wann er spricht.
Didaktisch ist das anschlussfähig. Ein Trainer, der nach jedem Item eingreift, erhöht die zu verarbeitende Informationsmenge genau dann, wenn die Aufgabe selbst noch Kapazität bindet. Die Simulation sieht ein Ausblenden der Unterstützung bereits vor, allerdings erst über die drei Fallvignetten hinweg. Der Vorschlag der Expertinnen verlegt dieses Fading auf eine feinere Ebene und ist im nächsten Zyklus zu prüfen. Er entspricht der Befundlage zu Reflexionsphasen, die nach Chernikova et al. (2020b) bei fortgeschrittenen Lernenden stärker wirken als bei Anfängerinnen und Anfängern.
Damit lässt sich der didaktische Teil der ersten Fragestellung beantworten. Die Simulation setzt die drei Ebenen diagnostischen Handelns aus Abschnitt 2.1 in einer Aufgabe zusammen: Die Studierenden führen die Elizitierung und die Hilfengebung selbst durch, verdichten Hilfe-Summe, V2- und SVK-Werte sowie Modifiability zu einem Urteil und begründen dieses im Debriefing. Die deterministische Antwortmatrix hält dabei den Fall über alle Durchläufe konstant und stellt über die Phasentrennung sicher, dass nur diagnostische Beiträge in die Wertung eingehen. Curricular verorten die Expertinnen das Werkzeug mehrheitlich nach der einführenden Vermittlung von MuSE-Pro und DA, ohne dass sich eine einheitliche Zielgruppe abzeichnet.

5.2 Interdisziplinäre Übertragbarkeit

Über die Sprachtherapie hinaus legt die Architektur eine Übertragung nahe. Die folgenden Überlegungen sind Designannahmen und Hypothesen für künftige Forschung, keine Ergebnisse der berichteten Studie; empirisch geprüft wurde ausschließlich der sprachtherapeutische Anwendungsfall. Plausibel erscheint eine Übertragung dort, wo ein standardisiertes Verfahren in einer

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

Untersuchungssituation zwischen zwei Personen realistisch und zugleich reproduzierbar nachgestellt werden soll, sodass alle Studierenden denselben Fall erleben.

Vier Gestaltungsprinzipien sind dabei inhaltsunabhängig formulierbar: (1) eine vorab festgelegte Antwortmatrix, die das Fallverhalten über alle Durchläufe konstant hält; (2) die funktionsgetrennte Verarbeitung der Lernendenbeiträge; (3) eine gemeinsame Datenschnittstelle, über die Oberfläche und System denselben Stand teilen; (4) die sichtbare wie hörbare Trennung der Sprecherrollen.
Auf dieser Grundlage ließe sich das System erweitern. Weitere Rollen wie eine Bezugsperson oder eine Kollegin könnten ergänzt werden, ohne dass die zugrunde liegende Logik geändert werden muss; ebenso ließen sich Bild- oder Videoimpulse einbinden (technische Umsetzung; vgl. 3.3). Damit ist die Bauweise offen für komplexere Untersuchungssituationen, ohne ihre Reproduzierbarkeit zu verlieren.
Als nächste Prüffelder bieten sich benachbarte Diagnostik-Domänen an, etwa die klinisch-psychologische Diagnostik, die medizinische Anamnese mit standardisierten Patient:innen oder die formative Lernstandsdiagnostik in der Lehrkräftebildung (Codreanu et al., 2020). Ob die vier Prinzipien dort tragen, ist offen: Jede dieser Domänen müsste zunächst zeigen, dass sich ihre diagnostisch relevanten Antworten ähnlich geschlossen klassifizieren lassen wie die Zielstrukturen des hier verwendeten Testverfahrens. Damit ist der technische Teil der ersten Fragestellung beantwortet.

5.3 Ethische und datenschutzrechtliche Erwägungen

Der Einsatz generativer KI in der Lehre wirft Fragen auf, die über die Einwilligung zur Aufzeichnung hinausgehen. Drei sind hier einschlägig. Die erste betrifft die Sprachdaten, da Sprachsynthese und Spracherkennung über Dienste außerhalb der Hochschule laufen. Eine Sprachaufnahme ist ein personenbezogenes Datum, den Status biometrischer Daten im Sinne von Art. 9 DSGVO erhält sie jedoch erst bei einer Verarbeitung zur Identifizierung, die hier nicht stattfindet. Da die Sprechenden gleichwohl identifizierbar bleiben, sind die Datenschutz-Grundverordnung und das Bayerische Datenschutzgesetz maßgeblich. Daher ist das Vorhaben vor dem Einsatz in der Lehre mit dem Datenschutzbeauftragten der LMU abzustimmen. Als Schutzmaßnahmen bleibt die Spracheingabe optional, wird serverseitig vermittelt und mit einer Löschfrist versehen. Kritisch bleibt allerdings, dass sich die Übermittlung an Dritte ohne ein hochschuleigenes Modell nicht vermeiden lässt.

Die zweite Frage betrifft die maschinelle Rekonstruktion von Kompetenzen und berührt beide Gruppen. Protokolliert werden Simulationsverlauf, Äußerungen und Urteile; in der Pilotierung bildet das System daraus zusätzlich ein Lernermodell, das den Umfang der Unterstützung bestimmt. Solche Daten sind besonders sensibel, weil sie fachliches Versagen einer identifizierbaren Person festhalten. Damit ist offen, ob eine Maschine Kompetenz überhaupt abbilden kann. Im strengen Sinn nicht: Diagnostische Kompetenz ist eine Disposition, beobachtbar ist allein ihre Anwendung (Heitzmann et al., 2019). Eine Hilfestufenfolge indiziert diesen Prozess, sie misst die Disposition nicht, beruht auf wenigen Minuten und ist psychometrisch ungeprüft. Möglich ist etwas Engeres: Weil die Kindantworten deterministisch festliegen, erhalten alle dasselbe Material, und Unterschiede im Vorgehen sind den eigenen Entscheidungen zuzurechnen. Dieser engere Anspruch eignet sich für formative Rückmeldung und Prozessforschung, nicht jedoch für die Leistungsbewertung. Aus diesen Überlegungen leiten sich konkrete Schutzmaßnahmen ab. Die Auswertung erfolgt ausschließlich pseudonymisiert, wobei das Lernermodell allein den Umfang der Unterstützung innerhalb der Sitzung steuert und nicht der Bewertung der Person dient. Ergänzend erhalten die Lernenden ein Einsichtsrecht in die über sie erhobenen Daten, für die zugleich eine feste Löschfrist gilt.

Die dritte Frage betrifft die Transparenz über die KI-gestützte Funktionsweise. Die Transparenzpflicht reicht dabei über den bloßen Hinweis auf die Beteiligung eines Sprachmodells hinaus. Für die Studierenden muss zum einen erkennbar sein, welche Teile der Sitzung festgelegt und welche generativ erzeugt sind und dass die Rückmeldungen des Meta-Trainers deshalb fachlich fehlerhaft sein können. Zum anderen ist offenzulegen, dass die drei Fälle konstruiert und nicht repräsentativ sind und dass jeder Turn protokolliert wird. Die vollständige Offenlegung ist jedoch nicht ohne Kosten. Wären Antwortmatrix und erwartete Modifiability während der Sitzung

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

einsehbar, ließe sich die Hilfengebung optimieren statt diagnostizieren, sodass die eigenständige Beobachtung der Veränderbarkeit als eigentliches Lernziel entfiele. Hinzu kommt, dass das Wissen um die Protokollierung genau das Verhalten verändern kann, das erfasst werden soll. Transparenz ist deshalb zu staffeln: Prinzipien und Datenflüsse werden vorab offengelegt, die einzelnen Festlegungen erst im Debriefing. Für die hier berichtete Expertinnenevaluation wurde kein Ethikvotum eingeholt, für die Pilotierung mit Studierenden ist dies jedoch vorgesehen.

5.4 Limitationen und Forschungsagenda

Die Studie ist in mehrfacher Hinsicht limitiert. Erstens erlaubt die Erhebung mit drei Expertinnen nur deskriptive Aussagen auf Individualebene. Mayring (2022) weist Klein-N-Studien allerdings ausdrücklich als legitimes Anwendungsgebiet der qualitativen Inhaltsanalyse aus. Zweitens sind alle drei Expertinnen aktuelle oder ehemalige Mitarbeiterinnen desselben Lehrstuhls, sodass die Urteile institutionell nicht unabhängig sind. Drittens erfolgte die Kodierung durch den Erstautor ohne quantifizierte Intercoderreliabilität, stattdessen wurden das Kategoriensystem und strittige Entscheidungen konsensuell im Autorenteam abgestimmt. Viertens lässt sich für Sitzung 1 eine akustische Mit-Wahrnehmung der parallel arbeitenden Expertinnen nicht ausschließen, wobei die Konvergenz mit Sitzung 2 jedoch gegen eine substanzielle Verzerrung spricht. Fünftens besteht die Stichprobe aus Hochschullehrenden, nicht aus Studierenden, sodass die Studie eine fachlich-didaktische Tauglichkeitseinschätzung liefert und keinen Lernwirksamkeitsnachweis. Sechstens wurde inhaltlich ausschließlich die Fallvignette „Emma» im höchsten Support-Level erprobt, während die Plausibilitätsbestätigung von „Noah» und „Liam» noch aussteht. Siebtens wurde in einer der drei Sitzungen der Simulationsverlauf nicht vollständig protokolliert, sodass die Verhaltensdaten nur für zwei Erprobungen vorliegen.

Diese Vorbehalte sind zugleich der Ausgangspunkt der weiteren Forschungsagenda, die den ersten DBR-Zyklus fortschreibt. Im nächsten Schritt wird die Bedienoberfläche überarbeitet, in der die Expertinnen den deutlichsten Verbesserungsbedarf verorteten. Zugleich wird die Trainerrückmeldung innerhalb eines Falls schrittweise zurückgenommen, statt nach jedem Item zu erfolgen. Anschließend werden die Fallprofile „Noah“ und „Liam“ auf Reliabilität, fachliche Plausibilität und didaktische Tragfähigkeit geprüft.

Darauf aufbauend soll die überarbeitete Umgebung mit Studierenden pilotiert und in einem Prä-Post-Design untersucht werden. Hierfür ist ein mehrdimensionaler Vortest vorbereitet, welcher fünf Komponenten umfasst: (1) Technologiebereitschaft über die TK-Subskala des TPACK-Inventars (Schmidt et al., 2009), (2) Vorwissen über einen Wissenstest, (3) diagnostisches Denken über die Kurzform des Diagnostic Thinking Inventory (Schäfer et al., 2019), (4) Selbsteinschätzung über ein Confidence-Rating und (5) diagnostisches Reasoning über eine offene Fallvignette nach St. Marie et al. (2021). Der Wissenstest und das Confidence-Rating speisen das Lernermodell, während die übrigen Komponenten gemeinsam mit ihrer Post-Erhebung der Operationalisierung des Kompetenzzuwachses dienen. Das Reasoning wird entlang der Facetten Justification, Disconfirmation und Transparency (Bauer et al., 2022) kodiert, diesmal doppelt und mit einem berichteten Übereinstimmungsmaß. Die kognitive Last (Klepsch et al., 2017) und die Simulationszufriedenheit (Levett-Jones et al., 2011) ergänzen die Post-Erhebung.

6 Fazit

KI-gestützte Simulationen erscheinen als vielversprechender Ansatz, um diagnostische Kompetenzen in der sprachtherapeutischen Hochschulausbildung gezielt zu fördern. Die explorative Expertinnenbewertung deutet darauf hin, dass die entwickelte Simulationsumgebung fachlich anschlussfähig und didaktisch sinnvoll ist, sofern die notwendigen Vorkenntnisse zu MuSE-Pro und Dynamic Assessment gegeben sind. Als Einsatzkontexte bestätigen die Expertinnen übereinstimmend die Prüfungsvorbereitung und die Vorbereitung auf reale Fallarbeit, während ihre Einschätzungen über die Zielgruppe auseinandergehen. Das besondere Potenzial liegt darin, komplexe diagnostische Entscheidungsprozesse wiederholbar, geschützt und feedbackgestützt trainierbar zu machen, und zwar an einem Verfahren wie dem Dynamic Assessment, das in der Praxis bislang wenig etabliert ist. Zugleich zeigt die Evaluation präzise, dass vor allem die Bedienoberfläche und die Frage der Eingriffshäufigkeit des Trainers nachzubessern sind. Ob die Umgebung tatsächlich zu einem messbaren Kompetenzzuwachs führt, muss die geplante Pilotie-

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)

rung mit Studierenden zeigen. Damit eröffnet die Simulation einen Weg für die evidenzorientierte Weiterentwicklung der sprachtherapeutischen Hochschullehre.

Literaturverzeichnis

American Speech-Language-Hearing Association. (2016). *Standards for accreditation of graduate education programs in audiology and speech-language pathology.* https://caa.asha.org/wp-content/uploads/Accreditation-Standards-for-Graduate-Programs.pdf.

Barab, S. (2006). Design-based research: A methodological toolkit for the learning scientist. In R. K. Sawyer (Hrsg.), *The Cambridge handbook of the learning sciences* (S. 153–169). Cambridge University Press.

Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., & Sailer, M. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education: An interdisciplinary comparison. *Frontiers in Psychology, 11,* Article 562665. https://doi.org/10.3389/fpsyg.2020.562665

Bauer, E., Sailer, M., Kiesewetter, J., Fischer, M. R., & Fischer, F. (2022). Diagnostic argumentation in teacher education: Making the case for justification, disconfirmation, and transparency. *Frontiers in Education, 7*, Article 977631. https://doi.org/10.3389/feduc.2022.977631

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology, 54*(5), 1222–1245. https://doi.org/10.1111/bjet.13336

Bauer, E., Heitzmann, N. Bannert, M., Chernikova, O., Fischer, M.R., Hofer, S., Holzberger, D., Kasneci, E., Koenen, J., Kosel , C., Küchemann, S., Michaeli, T., Neuhaus, B.J., Niklas, F., Obersteiner, A., Pfeffer, J., Sailer, M., Schmidmaier, R., Schmidt-Hertha, B., Stadler, M. Ufer, S., Vorholzer, A., Seidel, T., & Fischer, F.(2025). Personalizing simulation-based learning in higher education. *Learning and Individual Differences, 122,* Article 102746. https://doi.org/10.1016/j.lindif.2025.102746

Berg, M. (2015). *MuSE-Pro. Überprüfung grammatischer Fähigkeiten bei 5- bis 8-jährigen Kindern. Morphologische und Syntaktische Entwicklung – Produktion.* Reinhardt.

Beushausen, U., & Walther, W. (2010). Clinical Reasoning in der Logopädie. Therapeutische Entscheidungen bewusst treffen und fundiert begründen. *Forum Logopädie, 24*(4), 30–37.

Brooke, J. (1996). SUS – A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Hrsg.), *Usability evaluation in industry* (S. 189–194). Taylor & Francis.

Carter, M. D. (2019). The effects of computer-based simulations on speech-language pathology student performance. *Journal of Communication Disorders, 77,* 44–55. https://doi.org/10.1016/j.jcomdis.2018.12.006

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020a). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F. (2020b). Facilitating diagnostic competences in higher education: A meta-analysis in medical and teacher education. *Educational Psychology Review, 32,* 157–196. https://doi.org/10.1007/s10648-019-09492-2

Chernikova, O., Sommerhoff, D., Stadler, M., Holzberger, D., Nickl, M., Seidel, T., Kasneci, E., Küchemann, S., Kuhn, J., Fischer, F., & Heitzmann, N. (2025). Personalization through adaptivity or adaptability? A meta-analysis on simulation-based learning in higher education. *Educational Research Review, 46,* Article 100662. https://doi.org/10.1016/j.edurev.2024.100662

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2020). Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teaching and Teacher Education, 95,* Article 103146. https://doi.org/10.1016/j.tate.2020.103146

Delage, H., Prat, P., & Kehoe, M. (2021). Évaluation dynamique en orthophonie / logopédie. *Glossa, (131),* 121–142. https://doi.org/10.61989/cdz05v79

Diepeveen, S., Van Haaften, L., Terband, H., De Swart, B., & Maassen, B. (2020). Clinical reasoning for speech sound disorders: Diagnosis and intervention in speech-language pathologists' daily practice. *American Journal of Speech-Language Pathology, 29*(3), 1529–1549. https://doi.org/10.1044/2020_ajslp-19-00040

Dudding, C. C., & Pfeiffer, D. L. (2018). Clinical decision-making in speech-language pathology graduate students: Quantitative findings. *Teaching and Learning in Communication Sciences and Disorders, 2*(1), Article 2. https://doi.org/10.30707/tlcsd2.1dudding

Ehlert, H. (2021a). *Dynamic Assessment. Prozess und Potential in der Diagnostik von Sprachentwicklungsstörungen.* Springer VS.

Ehlert, H. (2021b). Dynamic Assessment: Potenzial in der Diagnostik mehrsprachiger Kinder. *Sprachförderung und Sprachtherapie in Schule und Praxis, 10*(2), 88–93.

Ehlert, H. (im Druck). Logopädische Diagnostik bei mehrsprachigen Kindern: Herausforderungen und Chancen am Beispiel von Dynamic Assessment. In K. Albrecht & C. Haupt (Hrsg.), *Sprachtherapeutische Diagnostik und Intervention bei mehrsprachigen Kindern.* Hogrefe.

ElevenLabs Inc. (2024). *ElevenLabs Voice AI.* https://elevenlabs.io

Eppich, W., & Cheng, A. (2015). Promoting Excellence and Reflective Learning in Simulation (PEARLS): Development and rationale for a blended approach to health care simulation debriefing. *Simulation in Healthcare, 10*(2), 106–115. https://doi.org/10.1097/SIH.0000000000000072

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45. https://doi.org/10.14786/flr.v2i3.96

Fischer, F., Bauer, E., Seidel, T., Schmidmaier, R., Radkowitsch, A., Neuhaus, B. J., Hofer, S. I., Sommerhoff, D., Ufer, S., Kuhn, J., Küchemann, S., Sailer, M., Koenen, J., Gartmeier, M., Berberat, P., Frenzel, A., Heitzmann, N., Holzberger, D., Pfeffer, J., … Fischer, M. R. (2022). Representational scaffolding in digital simulations – Learning professional practices in higher education. *Information and Learning Sciences, 123*(11/12), 645–665. https://doi.org/10.1108/ILS-06-2022-0076

Ford, C., McNally, D., & Ford, K. (2017). Using design-based research in higher education innovation. *Online Learning, 21*(3), 50–67. https://doi.org/10.24059/olj.v21i3.1232

Gama, A. C. C., Quinino, R. da C., Medeiros, A. M., Mancini, P. C., Mourão, A. M., Santos, L. G., Machado, T. H., & Gomes, N. R. (2025). FonoTCS: Validation of a tool for assessing clinical reasoning in speech-language pathology. *CoDAS, 37*(3), Article e20240206. https://doi.org/10.1590/2317-1782/e20240206en

Georgiou, G. P. (2025). Transforming speech-language pathology with AI: Opportunities, challenges, and ethical guidelines. *Healthcare, 13*(19), Article 2460. https://doi.org/10.3390/healthcare13192460

Hadjadj, O., Kehoe, M., & Delage, H. (2022). Dynamic Assessment identifies morphosyntactic deficits in mono- and bilingual children with developmental language disorder. *Languages, 7*(4), Article 295. https://doi.org/10.3390/languages7040295

Hasson, N., Dodd, B., & Botting, N. (2012). Dynamic Assessment of Sentence Structure (DASS): Design and evaluation of a novel procedure for the assessment of syntax in children with language impairments. *International Journal of Language & Communication Disorders, 47*(3), 285–299. https://doi.org/10.1111/j.1460-6984.2011.00108.x

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. F*rontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Hewat, S., Penman, A., Davidson, B., Baldac, S., Howells, S., Walters, J., Purcell, A., Cardell, E., McCabe, P., Caird, E., Ward, E., & Hill, A. E. (2020). A framework to support the development of quality simulation-based learning programmes in speech–language pathology. *International Journal of Language & Communication Disorders, 55*(2), 287–300. https://doi.org/10.1111/1460-6984.12515

Hill, A. E., Ward, E., Heard, R., McAllister, S., McCabe, P., Penman, A., Caird, E., Aldridge, D., Baldac, S., Cardell, E., Davenport, R., Davidson, B., Hewat, S., Howells, S., Purcell, A., & Walters, J. (2021). Simulation can replace part of speech-language pathology placement time: A randomised controlled trial. *International Journal of Speech-Language Pathology, 23*(1), 92–102. https://doi.org/10.1080/17549507.2020.1722238

Hoben, K., Varley, R., & Cox, R. (2007). Clinical reasoning skills of speech and language therapy students. International *Journal of Language & Communication Disorders, 42*(S1), 123–135. https://doi.org/10.1080/13682820601171530

Kauschke, C., Lüke, C., Dohmen, A., Haid, A., Leitinger, C., Männel, C., Penz, T., Sachse, S., Scharff Rethfeldt, W., Spranger, J., Vogt, S., Neumann, K., & Niederberger, M. (2023). Delphi-Studie zur Definition und Terminologie von Sprachentwicklungsstörungen: Eine interdisziplinäre Neubestimmung für den deutschsprachigen Raum. *Logos, 31*(2), 84–102.

Kiesewetter, J., Hege, I., Sailer, M., Bauer, E., Schulz, C., Platz, M., & Adler, M. (2022). Implementing remote collaboration in a virtual patient platform: Usability study. *JMIR Medical Education, 8*(3), Article e24306. https://doi.org/10.2196/24306

Klein, M., Otto, B., Fischer, M. R., & Stark, R. (2019). Fostering medical students' clinical reasoning by learning from errors in clinical case vignettes: Effects and conditions of additional prompting procedures to foster self-explanations. *Advances in Health Sciences Education, 24,* 331–351. https://doi.org/10.1007/s10459-018-09870-5

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8,* Article 1997. https://doi.org/10.3389/fpsyg.2017.01997

Levett-Jones, T., McCoy, M., Lapkin, S., Noble, D., Hoffman, K., Dempsey, J., Arthur, C., & Roche, J. (2011). The development and psychometric testing of the satisfaction with simulation experience scale. *Nurse Education Today, 31*(7), 705–710. https://doi.org/10.1016/j.nedt.2011.01.004

Li, K., Liu, T., Bashkansky, N., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2024). Measuring and controlling instruction (in)stability in language model dialogs. In *Proceedings of the First Conference on Language Modeling* (COLM 2024). https://arxiv.org/abs/2402.10962

Mayring, P. (2022). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (13., überarb. Aufl.). Beltz.

Nagdee, N., Sebothoma, B., Madahana, M., Khoza-Shangase, K., & Moroe, N. (2022). Simulations as a mode of clinical training in healthcare professions: A scoping review to guide planning in speech-language pathology and audiology during the COVID-19 pandemic and beyond. *South African Journal of Communication Disorders, 69*(2), Article a905. https://doi.org/10.4102/sajcd.v69i2.905

OpenAI. (2022). *Introducing Whisper.* https://openai.com/index/whisper/

Reinmann, G. (2005). Innovation ohne Forschung? Ein Plädoyer für den Design-Based Research-Ansatz in der Lehr-Lernforschung. *Unterrichtswissenschaft, 33*(1), 52–69.

Saccone, C., Murray, J., Gunasekaran, S., & Doeltgen, S. H. (2025). Clinical reasoning and decision-making of pre-professional speech-language therapists during clinical swallowing examination. *International Journal of Language & Communication Disorders, 60*(3), Article e70028. https://doi.org/10.1111/1460-6984.70028

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83,* Article 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Schäfer, A. G. M., Sebelin, B., & Spitzer, L. (2019). Cultural adaption and validation of the German version of the Diagnostic Thinking Inventory (DTI-G). *International Journal of Health Professions, 6*(1), 32–45. https://doi.org/10.2478/ijhp-2019-0002

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological Pedagogical Content Knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education, 42*(2), 123–149. https://doi.org/10.1080/15391523.2009.10782544

Schwob, S., Tillé, Y., & Skoruppa, K. (2025). A comparison of two dynamic assessment situations for detecting development language disorder in monolingual and bilingual children. *Clinical Linguistics & Phonetics, 39*(6–8), 540–558. https://doi.org/10.1080/02699206.2024.2435010

Spreer, M. (2018). *Diagnostik von Sprach- und Kommunikationsstörungen im Kindesalter. Methoden und Verfahren.* Reinhardt.

St. Marie, B., Jimmerson, A., Perkhounkova, Y., & Herr, K. (2021). Developing and establishing content validity of vignettes for healthcare education and research. *Western Journal of Nursing Research, 43*(7), 677–685. https://doi.org/10.1177/0193945920969693

Suh, H., Dangol, A., Meadan, H., Miller, C. A., & Kientz, J. A. (2024). Opportunities and challenges for AI-based support for speech-language pathologists. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24).* Association for Computing Machinery. https://doi.org/10.1145/3663384.3663387

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36(*1), 157–178. https://doi.org/10.2307/41410412

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Hrsg.). Harvard University Press.

Zu den Autoren

Johannes Winterstein ist Student im BA Sprachtherapie an der Ludwig-Maximilians-Universität (LMU) München; Bachelorarbeit zum Thema „KI-gestützte Trainingssimulation (SHARP Frameworks) für Sprachtherapie-Studierende zum Dynamic Assessment der Verbzweitstellung".

Dr. Maximilian Hamann ist Akademischer Oberrat am Lehrstuhl für Sprachheilpädagogik der Ludwig-Maximilians-Universität (LMU) München.

Korrespondenzadresse

Johannes.Winterstein@campus.lmu.de
Max.Hamann@edu.lmu.de

 – Urheberrechtlich geschütztes Material – © Deutsche Gesellschaft für Sprachheilpädagogik e.V. (dgs)