

Zeitschrift für Fremdsprachenforschung 2025

Joel Guttke & Thorsten Merse

Der Theoriebegriff in der fremdsprachendidaktischen Forschung

Meral Roeben, Christiane Lütge, Florian Schultz-Pernice, Johanna Vejvoda, Michael Sailer,

Frank Fischer & Nicole Heitzmann

Diagnosing Cognitive Engagement in TEFL

Gabriela Lüthi, Elisabeth Peyer & Nadia Ravazzini

Speaking Tests in the Lower Secondary Foreign Language Classroom

Zeitschrift für Fremdsprachenforschung

Organ der Deutschen Gesellschaft
für Fremdsprachenforschung

Die ZFF wird herausgegeben von:

Mark Bechtel (Osnabrück)

Sophie Engelen (Gießen)

Andreas Grünewald (Bremen)

Henning Rossa (Trier)

Julia Settinieri (Bielefeld)

Eva Wilden (Duisburg-Essen)

Katrin Wisniewski (Leipzig)

ZFF – Zeitschrift für Fremdsprachenforschung

Band 36, Heft 2/2025

Redaktioneller Beirat:

Prof. Dr. Dagmar Abendroth-Timmer (Siegen)
Prof. Dr. Eva Burwitz-Melzer (Gießen)
Prof. Dr. Daniela Caspari (Berlin)
Prof. Dr. Britta Hufeisen (Darmstadt)
Prof. Dr. Friederike Klippel (München)
Prof. Dr. Grit Mehlhorn (Leipzig)
Prof. Dr. Jürgen Kurtz (Gießen)
Prof. Dr. Hélène Martinez (Gießen)
Prof. Dr. Nicole Marx (Köln)
Prof. Dr. Michael Legutke (Gießen)
Prof. Dr. Claudia Riemer (Bielefeld)
Prof. Dr. Dominik Rumlich (Paderborn)
Prof. Dr. Michael Schart (Jena)
Prof. Dr. Lars Schmelter (Wuppertal)
Prof. Dr. Barbara Schmenk (Waterloo)

Anschrift der Redaktion (verantwortlich):

Redaktion ZFF
c/o Prof. Dr. Andreas Grünewald
Universität Bremen – Fachbereich 10
Bibliotheksstr. 1
28359 Bremen
zff@dgff.de

Technische Redaktion:

Andreas Grünewald (Bremen)
Megan Dwinger (Bremen)

Anzeigen: wbv Media GmbH & Co. KG, Cira Korfmacher, Auf dem Esch 4, 33619 Bielefeld
Tel. (05 21) 9 11 01-251, E-Mail: cira.korfmacher@wbv.de

Erscheinungsweise: Jährlich 2 Ausgaben

2025 wbv Publikation
ein Geschäftsbereich der
wbv Media GmbH & Co. KG, Bielefeld

Gesamtherstellung:
wbv Media, Bielefeld,
Auf dem Esch 4, 33619 Bielefeld, ser
vice@wbv.de
wbv.de

eISSN: 3052-5411
DOI: 10.3278/ZFF2502W

Printed in Germany

Diese Publikation ist frei verfügbar zum Download
unter wbv-open-access.de

Diese Publikation ist unter folgender Creative-
Commons-Lizenz veröffentlicht:
<http://creativecommons.org/licenses/by-sa/4.0/>



Für alle in diesem Werk verwendeten Warennamen
sowie Firmen- und Markenbezeichnungen können
Schutzrechte bestehen, auch wenn diese nicht als
solche gekennzeichnet sind. Deren Verwendung in
diesem Werk berechtigt nicht zu der Annahme, dass
diese frei verfügbar seien.

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie;
detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Inhalt

Joel Guttke & Thorsten Merse

Der Theoriebegriff in der fremdsprachendidaktischen Forschung 5

Meral Roeben, Christiane Lütge, Florian Schultz-Pernice, Johanna Vejvoda, Michael Sailer, Frank Fischer & Nicole Heitzmann

Diagnosing Cognitive Engagement in TEFL 27

Gabriela Lüthi, Elisabeth Peyer & Nadia Ravazzini

Speaking Tests in the Lower Secondary Foreign Language Classroom 49

Der Theoriebegriff in der fremdsprachendidaktischen Forschung

Ein Systematisierungsvorschlag

JOEL GUTTKE¹ & THORSTEN MERSE²

Abstract

Advancing theory is a major concern in research into foreign language education. Despite its relevance, theory remains an uncontested issue in foreign language education. Drawing on frameworks from the Humanities and Social Sciences, this article proposes a heuristic of theory in foreign language education that spans three dimensions: What is the nature of theory? What does *theory* refer to? What is the purpose of theory? These three approaches to the concept of theory are illustrated using studies from the field of foreign language education. The heuristic bears the potential to structure theory work in foreign language education and to clarify the theory discourse between advocates of conceptual and empirical research into foreign language education.

1 Einleitung: Theorie! Theorie?

Für die Fremdsprachendidaktik sind Theorien von zentraler Bedeutung. In forschungsmethodischen Publikationen wird in Abgrenzung zu historischer und empirischer Forschung mit theoretisch-konzeptioneller Forschung ein eigenes Paradigma identifiziert, das insbesondere die Theorie- und Modellbildung fokussiert (Reimann, 2020; Schmenk, 2022a). Doch auch empirische Arbeiten fußen auf Theorien und trügen durch den Einsatz hypothesengenerierender und -prüfender Verfahren zur Theoriebildung bei (Schramm, 2021; Caspari, 2022), sodass die Auseinandersetzung mit Theoriebildung nicht ausschließlich der theoretisch-konzeptionellen Forschung vorbehalten ist. Vielmehr ist Theoriearbeit grundlegend in dem Erkenntnisinteresse jeder fremdsprachendidaktischen Forschung angelegt. Lehrlernprozesse im Fremdsprachenunterricht lassen sich „durch das forschungsgeleitete Aufstellen, empirische Überprüfen und erkenntnisbasierte Ausschärfen von *theoretischen Grundlagen, Begriffen, Konzepten und Modellen*“ (Caspari, 2022, S. 11; Hervorhebung der Verfasser) nur dann verstehen und optimieren, wenn Forscher:innen die dafür notwendigen theoretischen Konstrukte wie Begriffe,

1 Dr. Joel Guttke, Universität Duisburg-Essen, Institut für Anglophone Studien, Didaktik des Englischen, Universitätsstraße 12, 45141 Essen, E-Mail: joel.guttke@uni-due.de

2 Prof. Dr. Thorsten Merse, Universität Duisburg-Essen, Institut für Anglophone Studien, Didaktik des Englischen, Universitätsstraße 12, 45141 Essen, E-Mail: thorsten.merse@uni-due.de

Konzepte und Modelle (weiter-)entwickeln (z. B. zum Themenkomplex des kulturellen Lernens; König et al., 2022). Beide Argumente deuten darauf hin, dass Forschung zu Fremdsprachenunterricht immer auch Theorien erfordert, einfordert und mit produziert.

Trotz ihrer prominenten Stellung wird *Theorie* in der Fremdsprachendidaktik – sogar in Publikationen, die sich explizit mit Forschungsmethodologie befassen – kaum expliziert. 2025 hält der Theoriebegriff erstmals Einzug in das Metzler Lexikon Fremdsprachendidaktik. Darin attestieren Legutke und Schart (2025) dem Theoriebegriff „einen hohen Grad an Ambiguität“ (S. 472), der daraus resultiere, dass Theorie „gemeinlich als eine Sammelbezeichnung für alle Ergebnisse von wissenschaftlicher Tätigkeit benutzt“ (S. 472) werde. Ansonsten erfolgt(e) die Annäherung an Theoriearbeit bisher indirekt, indem z. B. hinsichtlich theoretisch-konzeptioneller Forschung Referenzarbeiten präsentiert (Schramm, 2021), Funktionen beschrieben (Legutke, 2022), oder forschungspraktische Empfehlungen formuliert wurden (Schmenk, 2022a). In welchem Verhältnis diese Aspekte zum Theoriebegriff stehen, wird hingegen nicht adressiert. Anders als in ihren Bezugsdisziplinen (z. B. der Zweitspracherwerbsforschung; VanPattem et al., 2020) bleibt in der Fremdsprachendidaktik folglich wegen fehlender oder unzureichender Explikation unklar, mit welchem Theoriebegriff operiert wird. Vor dem Hintergrund dieser inhaltlichen Unschärfe ist es zumindest fragwürdig, inwiefern der Begriff seine Funktion als Bezugspunkt fremdsprachendidaktischer Forschungsarbeit einzulösen vermag.

Der vorliegende Aufsatz hat zum Ziel, einen Beitrag zur Präzisierung fremdsprachendidaktischer Theoriearbeit zu leisten. Dazu wird basierend auf bereits publizierten Systematisierungsansätzen eine Heuristik des Theoriebegriffs entwickelt und exemplarisch anhand fremdsprachendidaktischer Forschungsarbeiten illustriert. Unter Heuristik werden hier in Anlehnung an Bender und Müller (2020) epistemische Vereinfachungs- und Annäherungsoperationen verstanden, um eine komplexitätsreduzierende Orientierung in die durchaus komplexe Fragestellung zu bringen, was in der Fremdsprachendidaktik mit *Theorie* gemeint ist. Die Heuristik bietet sich folglich als ein reflektiertes „Such- und Findeverfahren“ (Kleining, 2020, S. 4) an, das zur Genese neuer Erkenntnisse für das eigene Fach im Kontext der Theoriearbeit angewendet werden kann. Wir entwickeln den Aufsatz in folgendem Dreischritt:

- Erstens wird der fremdsprachendidaktische Diskurs zum Theoriebegriff nachgezeichnet, um aufzuzeigen, dass *Theorie* bisher selten Explikation erfährt (Abschnitt 2).
- Zweitens werden unter Bezugnahme auf sozial- und kulturwissenschaftliche Arbeiten von Abend (2008), Sandberg und Alvesson (2021) und Zima (2017) drei Bedeutungsebenen identifiziert, anhand derer sich die Ambiguität des Theoriebegriffs (Legutke & Schart, 2025) differenziert beschreiben lässt. Diese Bedeutungsebenen werden in einer Heuristik veranschaulicht (Abschnitt 3) und anschließend in fremdsprachendidaktischen Referenzarbeiten exemplarisch herausgearbeitet, wodurch die Vielseitigkeit von Theorieverständnissen greifbar wird (Abschnitt 4).

- Drittens werden die Potenziale und Herausforderungen der Heuristik für die Weiterentwicklung der fremdsprachendidaktischen Theoriearbeit reflektiert (Abschnitt 5).

Die Präsentation und Anwendung der Heuristik wird darlegen, dass die Bezeichnung unterschiedlicher Formen der Theoriearbeit durch den Sammelbegriff *Theorie* zu der impliziten Fehlannahme eines geteilten Theorieverständnisses führt. Das Potenzial des Beitrags besteht somit in der Bereitstellung einer Heuristik, die zukünftig zur expliziten Klärung des Theoriebegriffs genutzt werden kann, um den Theoriediskurs in der Fremdsprachendidaktik zu präzisieren.

2 *Theorie* im forschungsmethodologischen Diskurs der Fremdsprachendidaktik

In einem ersten Schritt untersuchen wir die bisherige Auseinandersetzung mit Theorie in der Fremdsprachendidaktik. Dazu beziehen wir uns auf Handbücher und Lexika, die als Grundlagenliteratur forschungsmethodische wie -methodologische Überlegungen der Disziplin kondensieren.

Den bislang einzigen Vorschlag einer intensionalen³ Theoriedefinition für die Fremdsprachendidaktik wagen Legutke und Schart (2025). Unter Verweis auf Abend (2008), einem der zentralen Bezugspunkte des vorliegenden Aufsatzes, unterscheiden sie zwischen Theorien im engeren und im weiteren Sinne. Theorien im engeren Sinne entsprechen „Aussagensysteme[n], die beobachtbare Phänomene [...] und ihre Einflussfaktoren beschreiben bzw. einheitliche Erklärungen zu verschiedenen Phänomenen liefern“ (Legutke & Schart, 2025, S. 472). Theorien im weiteren Sinne hingegen sind „nicht darauf angewiesen, sich in einem Prozess der empirischen Überprüfung zu bewähren. Als normative Setzungen, heuristische Modelle oder methodologische Überlegungen treiben sie den wissenschaftlichen Austausch [z. B. durch Systematisierung oder Umdeutung] voran“ (Legutke & Schart, 2025, S. 473). Entlang dieser Zweiteilung weisen Legutke und Schart (2025) Theorien im engeren bzw. weiteren Sinne eindeutig dem empirischen bzw. theoretisch-konzeptionellen Forschungsparadigma zu, betonen gleichzeitig aber auch die Interdependenz beider Lesarten des Theoriebegriffs. Der Mehrwert von Legutkes und Scharts (2025) Theoriedefinition besteht darin, dass die Autoren die Defizite des Theoriebegriffs, wie er bislang in der Fremdsprachendidaktik verwendet wurde, klar benennen und eine erste Begriffsdifferenzierung vornehmen. Wie nachfolgend zu zeigen sein wird, greift der Rückbezug auf die empirische Falsifizierbarkeit als Unterscheidungskriterium für Theorien im engeren und

3 Die intensionale und extensionale Definition beschreiben zwei Formen der Nominaldefinition, also der Explikation eines „zu definierende[n] Begriff[s] (Definiendum) durch einen oder mehrere bereits bekannte Begriffe (Definiens)“ (Döring, 2023, S. 226–227). Im Gegensatz zur extensionalen Nominaldefinition, bei der ein Begriff durch Beispiele beschrieben wird, enthält die intensionale Nominaldefinition Eigenschaften, über die sich ein Begriff charakterisieren lässt.

weiteren Sinne jedoch zu kurz und birgt die Gefahr, die Ambiguität des Theoriebegriffs auf den Dualismus von Empirie und Theorie zu reduzieren.

Darüber hinaus nähert man sich dem Theoriebegriff seitens der theoretisch-konzeptionellen Fremdsprachenforschung, die innerhalb der Disziplin auf eine lange Tradition zurückblickt (Reimann, 2020; Legutke, 2022) und sich *per definitionem* über den Theoriebegriff konstituiert:

- Bonnet (2017), Reimann (2020) und Volkmann (2022) verweisen auf die Hermeneutik als zentrales Verfahren der Erkenntnisgenerierung theoretisch-konzeptioneller Arbeiten und bezeichnen damit eine „beinahe unüberschaubare Methodenpluralität“ (Bonnet, 2017, S. 84). Gemeint sind damit vermutlich unterschiedliche Formen der Textarbeit, von denen alle Autoren eine Auswahl benennen (z. B. strukturelle, komparative, experimentelle Interpretation; Volkmann, 2022, S. 242).
- Schramm (2021) präsentiert Referenzarbeiten und illustriert daran Operationen theoretisch-konzeptioneller Forschung. Dazu zählt der Vergleich von Konzepten über verschiedene Disziplinen hinweg, das „Nachdenken über didaktisch-methodische Implikationen“ von Modellen und die „gewinnbringende Zusammenschau empirischer Studien“ (S. 216). Einen ähnlichen Ansatz verfolgt Legutke (2022), der theoretisch-konzeptionelle Forschungsarbeiten anhand ihrer Funktion zu ordnen versucht. Diesbezüglich unterscheidet er sieben Funktionen: „Entwicklung umfassender Konzepte der Sprachvermittlung, Entwicklung und/oder kritische Analyse tragender Konstrukte der Fremdsprachendidaktik, Modellbildung, Analyse und Auswahl von Lehr- und Lernmaterial, phänomenologische Arbeiten, bildungswissenschaftliche und bildungspolitische Positionierungen, vergleichende Überblicksforschungen“ (S. 41–44). Auch wenn die Aufzählungen als nicht erschöpfend kommentiert werden, zeigen sich zwischen den Autor:innen große Überschneidungen, aber auch einzelne Widersprüche. Während Legutke (2022) die Analyse und Auswahl von Lehr- und Lernmaterial als einen Prototyp theoretisch-konzeptioneller Forschung klassifiziert, ordnet Schramm (2021) sie den qualitativ-empirischen Forschungsarbeiten zu.
- Am stärksten konkretisiert Schmenk (2022a) theoretisch-konzeptionelle Forschung, indem sie Qualitätsmerkmale für die Auswahl von Texten als Grundlage für Theoriearbeit formuliert. Diese Merkmale weisen zumindest in Teilen Parallelen zu Volkmanns (2022) Impulsen auf, die er zur selbstkritischen Reflexion der Theoriearbeit formuliert.

Obwohl der Theoriebegriff nicht explizit bestimmt wird, sondern stattdessen vor allem Operationen der Theoriearbeit in den Fokus rücken, ist den Beiträgen gemeinsam, dass sie Texten als Informationsquelle sowie der hermeneutischen Untersuchung von Texten zentralen Wert für die Theoriearbeit zuschreiben. Die mehrheitlich geteilten Annahmen der zitierten Autor:innen lassen auf ein gemeinsames – wenn auch implizites – Theorieverständnis schließen, das wir in Abschnitt 3 (dort als Theorie₄) erneut aufgreifen. Zusammenfassend sei festgehalten, dass die gesichteten Publikationen den Theoriebegriff in der Fremdsprachendidaktik unzureichend charakterisieren und

die Disziplin bisher kaum handlungsleitende oder auch facettenreiche Charakterisierungen des Theoriebegriffs explizit ausgehandelt hat. Dennoch sind die in diesem Abschnitt präsentierten Charakteristika der Theoriearbeit anschlussfähig an Diskurse zum Theoriebegriff aus benachbarten Disziplinen. Diese Diskurse zeichnen wir nachfolgend nach und spannen daran unsere Heuristik zur Systematisierung des Theoriebegriffs in der Fremdsprachendidaktik auf.

3 Bedeutungsdimensionen des Theoriebegriffs

Die Frage nach dem Inhalt des Theoriebegriffs lässt sich entlang unterschiedlicher Begriffsdimensionen diskutieren. Abend (2008, S. 174) unterscheidet vier Dimensionen: die ontologische (Was ist eine Theorie?), evaluative (Was zeichnet eine *gute* Theorie aus?), teleologische (Was ist der Zweck einer Theorie?) und semantische (Was bedeutet *Theorie*?) Dimension. Gewiss sind diese Dimensionen nicht disjunkt, sondern bedingen sich gegenseitig. Die Qualität einer Theorie ist z. B. nur in Abhängigkeit eines konkreten semantischen Theorieverständnisses und mit Blick auf ihren Zweck beurteilbar. Wegen der sich hier bereits abzeichnenden Multidimensionalität des Theoriebegriffs geht mit jeder Aktivierung von *Theorie* in fremdsprachendidaktischen Forschungsdiskursen immer auch die Gefahr einher, dass Forschende gewissermaßen aneinander vorbeireden – vor allem dann, wenn *a priori* eine Verständigung über den Theoriebegriff fehlt.

Im Folgenden entwickeln wir nun eine Heuristik zur Systematisierung des Theoriebegriffs, um solche Verständigungen und Aushandlungen zu erleichtern. Dazu bedienen wir uns drei der von Abend (2008) aufgeworfenen Theoriedimensionen (ontologisch, teleologisch, semantisch) und bündeln daran Diskurse zum Theoriebegriff aus der Philosophie sowie den Geistes- und Sozialwissenschaften (vgl. Abbildung 1). Die Abbildung zeigt einen dreidimensionalen Raum, in dem sich fremdsprachendidaktische Forschungsarbeiten entsprechend des Theorieverständnisses, mit dem sie operieren, verorten lassen. Der Raum wird durch drei Achsen aufgespannt, welche die Dimensionen des Theoriebegriffs repräsentieren:

- Nach hinten gerichtete Achse/ontologische Dimension: Wie sind Theorien beschaffen? Je nach Beschaffenheit der Forschungsgegenstände, auf die sie sich beziehen, lassen sich Theorien anhand ihres *Formalisierungsgrades* entlang eines Kontinuums anordnen (Zima, 2017).
- Horizontale Achse/semantische Dimension: Was bezeichnet *Theorie*? Abend (2008) unterscheidet als Replik auf diese Frage sieben *Bedeutungen* des Theoriebegriffs.
- Vertikale Achse/teleologische Dimension: Wozu dient Theorie? Sandberg & Alvesson (2021) identifizieren fünf *Funktionen* des Theoriebegriffs.

Die Verortung von Forschungsarbeiten in dieser Heuristik kann nur idealtypisch sein, da wir andere Merkmalskombinationen grundsätzlich nicht ausschließen, wenngleich

einige Kombinationen vermutlich deutlicher wahrscheinlicher auftreten als andere. Es handelt sich hierbei um einen ersten Systematisierungsvorschlag, der keinen Anspruch auf Vollständigkeit erhebt. Die drei Dimensionen der Heuristik sind offen und durch den disziplinären Diskurs zu Theoriearbeit zu ergänzen. Die evaluative Dimension berücksichtigen wir in unserer Systematik bewusst nicht, da die Frage nach den Qualitätskriterien von Theorien – wie oben bereits angedeutet – unmittelbar durch das zugrundeliegende Theorieverständnis bedingt ist.

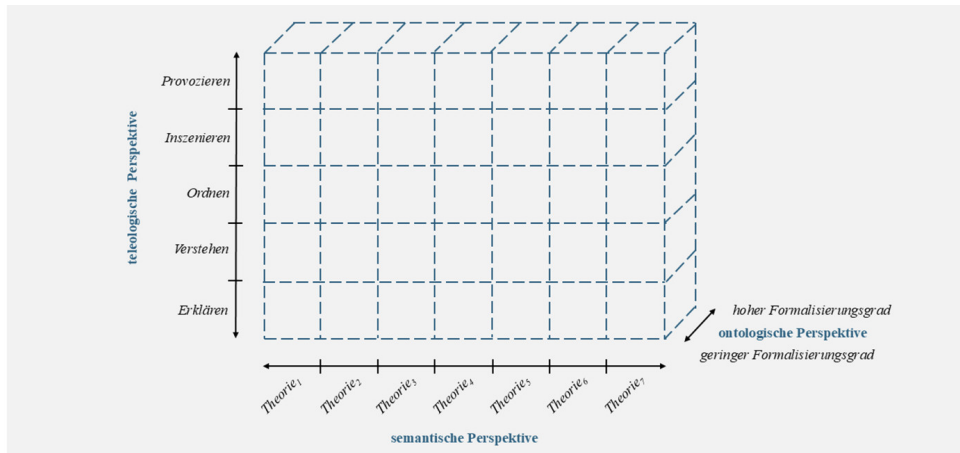


Abbildung 1: Systematik zur Klassifikation des Theoriebegriffs (Quelle: eigene Darstellung)

3.1 Ontologische Dimension: Konkurrierende wissenschaftstheoretische Annahmen zu *Theorie*

Beginnend mit der ontologischen Dimension stellen wir mit dem Kritischen Rationalismus (Gadenne, 2017) und der Dialogischen Theorie, die sich laut Zima (2017) auf die Kritische Theorie stützt, sie weiterentwickelt und neu deutet, zwei Ansätze einander gegenüber, um zu erörtern, inwiefern die Beschaffenheit der Forschungsgegenstände einer Disziplin Auswirkungen darauf hat, was eine Theorie konstituiert.

Als eines der wohl am stärksten mit dem Positivismusstreit⁴ assoziierten Werke gilt Karl Poppers (1934) *Logik der Forschung*. Sie bildet den Grundstein für die wissenschaftstheoretische Lehre des Kritischen Rationalismus. Ausgehend von einer Kritik an der Induktion erarbeitet Popper in seiner *Logik* eine hypothetisch-deduktive Alternative. „Wissenschaftliche Theorien sind allgemeine Sätze“ (Popper, 1934, S. 31), also Aussagensysteme bestehend aus Definitionen, Axiomen und Korollarien. Theorien setzten sich folglich aus Hypothesen zusammen, die einer empirischen Prüfung unterzogen werden könnten und sollten: „Ein empirisch-wissenschaftliches System muss an der Erfahrung scheitern können“ (Popper, 1934, S. 31). Das von Popper postulierte Falsifika-

4 Der Positivismusstreit bezeichnet eine Kontroverse, die in der Mitte des 20. Jahrhunderts zwischen Vertreter:innen des Kritischen Rationalismus (vertreten durch Karl R. Popper) und Vertreter:innen der Kritischen Theorie (vertreten durch Theodor W. Adorno) ausgetragen wurde. Gegenstand dieses Streits war die Frage nach adäquaten Methoden – und den damit verbundenen methodologischen Annahmen – zur Erforschung sozialwissenschaftlicher Fragestellungen.

tionsprinzip, das auch Legutke und Schart (2025) in ihrer Definition von Theorien im engeren Sinne aufgreifen, wird damit zu einem zentralen Abgrenzungskriterium wissenschaftlicher Theorien. Gleichzeitig erfordert es implizit die intersubjektive Nachvollziehbarkeit als Qualitätskriterium wissenschaftlicher Sätze. Vor diesem Hintergrund zeichnen sich Theorien durch einen hohen Grad an Formalisierung aus mit dem Ziel, Sachverhalte zu erklären oder zu prognostizieren (Keuth, 2017). Die Nähe von Poppers Ausführungen zu den Formal- und Naturwissenschaften ist unverkennbar. In diesem Sinn hoch formalisierte und zweckdienliche Theorien werden auch als *szientifisch* bezeichnet (Harant & Thomas, 2020). Intuitiv mag man nun die fremdsprachendidaktische Forschung mit ihrer geisteswissenschaftlichen Verortung möglicherweise eher weniger in dieser szientifischen Theorierichtung verorten. Aber auch in der Fremdsprachendidaktik gibt es vermehrt Arbeiten, deren Hypothesen oder theoretische Vorannahmen sich in mathematischen Modellen stark formalisieren lassen (z. B. Porsch et al., 2023).

Diametral gegenüber steht dem szientifischen Theorieverständnis Peter Zimas (2017) kultur- und sozialwissenschaftlich geprägter Entwurf einer Dialogischen Theorie. Ausgangspunkt dafür bildet die These, der Poppersche Theoriebegriff sei aufgrund seines radikalen Formalismus inkompatibel mit den Kultur- und Sozialwissenschaften, in denen Theorien aus Erzählungen und Diskursen bestünden, die sich kaum auf eine formallogische Struktur reduzieren ließen. Diese Inkompatibilität führt Zima (2017) auf die kulturelle, sprachliche und ideologische Bedingtheit von Theorien zurück:

Die Theorie ist ein von ideologischen Interessen geleiteter Diskurs, dessen Aussagesubjekt über seine Relevanzkriterien, seine semantisch-narrativen Verfahren und seine Akantenmodelle im sozio-linguistischen Kontext nachdenkt und sie als partikulare Konstruktionen einer ambivalenten, vieldeutigen Wirklichkeit auffaßt [sic!], deren Erkenntnis den Dialog mit anderen Theorien voraussetzt. (S. 62)

Diese dreifache Bedingtheit von Theorien wird nachführend knapp ausgeführt, um das Besondere von Theorien in den Kultur- und Sozialwissenschaften im Vergleich zu den Formal- und Naturwissenschaften herauszuarbeiten. Erstens resultiert die Herleitung der *kulturellen Bedingtheit* aus der Beobachtung, kultur- und sozialwissenschaftliche Theorien seien in ihrer Gültigkeit stets auf einen spezifischen sozio-kulturellen Kontext beschränkt: „Während etwa das Periodensystem der chemischen Elemente [...] überkulturelle Geltung beanspruchen kann, können die in Europa verwendeten kunst- oder literaturgeschichtlichen Periodisierungen keineswegs auf nichteuropäische Kulturen angewendet werden“ (Zima, 2017, S. 34). Diese Partikularität ergebe sich unmittelbar aus der Beschaffenheit der Gegenstände kultur- und sozialwissenschaftlicher Forschung. Theorien in diesen Disziplinen setzten sich mit Problemen sozialer Natur auseinander, deren Lösung „Engagement, Kritik und Wertung“ (Zima, 2017, S. 36) erfordere. Während die Kraft, die auf ein Objekt wirkt, physikalisch bestimmt werden kann, handelt es sich bei Mehrsprachigkeit beispielsweise um ein *von Menschen über Menschen sozial konstruiertes* Konzept, dessen Bedeutung und Bewertung zwischen gesellschaftlichen Kontexten variiert.

Zweitens träten kultur- und sozialwissenschaftliche Theorien in Form semantisch-narrativer Strukturen auf und seien folglich *sprachlich bedingt*. Zur Präzisierung der semantisch-narrativen Form führt Zima mit dem Soziolekt und den Aktantenmodellen zwei Diskurscharakteristika ein. „Der Soziolekt als ideologische, wissenschaftliche, literarische, philosophische oder religiöse Gruppensprache ist hier lediglich ein theoretisches Konstrukt, das verschiedene Diskurse [...] aufgrund ihrer gemeinsamen lexikalischen, semantischen und syntaktischen Merkmale bündelt“ (Zima, 2017, S. 52). Über diese konkret-sprachliche Ebene hinaus beschreibe das Aktantenmodell zentrale Charaktere und Handlungselemente von Diskursen. Die Art und Weise, wie eine Theorie versprachlicht wird, erlaube ihren Verfasser:innen folglich, sich in einem spezifischen Diskurs zu verorten. Rassismuskritische Arbeiten der Fremdsprachendidaktik rekurrieren zum Beispiel häufig auf bestimmte lexikalische Elemente (z. B. Macht, Dekolonisierung, *white privilege*) und etablierte Narrative (z. B. der Westen als unterdrückender Aktant). Die darin artikulierten Theorien lassen sich – anders als szientifische Theorien – nicht auf ein formallogisches Aussagensystem reduzieren, da sie erst durch sprachliche Markierungen einem gemeinsamen Diskurs zugeordnet werden können.

Aus diesen Ausführungen folge in gewissem Maße die *ideologische Bedingtheit* von Theorien, wobei der Ideologiebegriff bei Zima keinesfalls negativ konnotiert ist: „Alle theoretischen Soziolekte und ihre Diskurse sind ideologisch in dem Sinne, daß [sic!] sie Gruppeninteressen artikulieren, die im Diskurs auf bestimmte Probleme der sozilinguistischen Problematik reagieren“ (Zima, 2017, S. 56). Ob sich ein Individuum an einem theoretischen Diskurs beteiligt und wie diese Beteiligung sprachlich erfolgt, werde durch sein ideologisches Engagement beeinflusst. Um dem eigenen ideologischen Engagement nicht im Sinne eines blinden Flecks zu unterliegen und es unter Umständen sogar zu relativieren, betont Zima (2017, S. 62) die Orientierung an Dialog und Reflexion durch die fortlaufende Auseinandersetzung mit anderen Theorien, die er in seinem Werk selbst mehrfach demonstriert.

Dies impliziert auch für die Fremdsprachendidaktik, dass somit ein *Motor* für neue Forschungsarbeiten aktivierbar ist, der die Entwicklung von Theorien wider den Stillstand in Fluss hält. Exemplarisch lässt sich dies am Diskurs zum kulturellen Lernen verdeutlichen, in dem die Kritik an landeskundlichen Ansätzen unter Rückgriff auf neue Bezugstheorien zur Entwicklung von Alternativen (z. B. in Form von Inter- und Transkulturalität) geführt hat. Damit lässt sich auch einer erstarrenden „*sloganzation*“ (Schmenk et al., 2019, S. 4; Hervorhebung im Original) von Theorien in der Fremdsprachendidaktik vorbeugen, da konstante Reflexion und diskursive Kritik zentral in Zimas Theorieverständnis angelegt sind.

Die Gegenüberstellung von Popper und Zima verdeutlicht, dass die Frage danach, was Theorie ist (formallogisches Aussagensystem bis hin zu Narration), durch die Beschaffenheit der Forschungsgegenstände einer Disziplin (Objekte einer physikalischen Realität bis hin zu sozialen Konstrukten) und die daraus resultierenden wissenstheoretischen Annahmen unterschiedlich beantwortet werden muss. Die beiden Positionen stehen prototypisch für die Pole eines Kontinuums, das sich zwischen Extrema wie Universalität und Partikularität, Objektivität und Subjektivität sowie Formalismus und Dis-

kursivität erstreckt. Wie anhand der exemplarischen Forschungsarbeiten in Abschnitt 4 zu zeigen sein wird, lassen sich manche fremdsprachendidaktischen Theorien nicht im binären Sinne einem dieser Extrema zuordnen, sondern erfordern qualitative Abstufungen entlang des Kontinuums. Andererseits deutet sich an, dass die durch das Kontinuum repräsentierten Theorieformen unterschiedlich adäquat für den Einsatz über Teildisziplinen der Fremdsprachendidaktik hinweg erscheinen, die aufgrund der Beschaffenheit ihrer Forschungsgegenstände gezwungen sind, unter Bezugnahme auf unterschiedliche wissenstheoretische Annahmen zu operieren. Daraus folgt, dass in der Fremdsprachendidaktik grundsätzlich eine (überaus produktive!) Vielfalt von Theorien aktiv sein kann, es dann aber explizite Reflexion und Verständigung darüber braucht, mit welcher Theorieform – bedingt durch den jeweils eigenen Forschungsgegenstand – operiert wird.

3.2 Semantische Dimension: Was bezeichnet *Theorie*?

Wie im vorangegangenen Abschnitt deutlich wurde, könnten Poppers und Zimas wissenschaftstheoretische Grundannahmen, auf die sie sich zur Definition von Theorien beziehen, unterschiedlicher nicht sein. Dennoch verwenden beide *Theorie* zur Bezeichnung ihrer gedanklichen Konstrukte. Da die Beziehung zwischen dem *signifier* und dem *signified* somit nicht eindeutig ist, besteht die Gefahr von Missverständnissen immer dann, wenn der Theoriebegriff in Diskussionen nicht expliziert wird: „My argument is that in the case of ‘theory’ the problems stem from the erroneous belief that there is something – indeed, one thing – out there for the word ‘theory’ to really correspond to“ (Abend, 2008, S. 182). Zur Veranschaulichung seines Arguments unterscheidet Abend (2008) auf Grundlage soziologischer Forschungsarbeiten sieben Theoriebegriffe, die er mit Theorie₁₋₇ bezeichnet und die nachfolgend zusammengefasst werden.

Theorie₁: Vertreter:innen des ersten Theoriebegriffs beschreiben mit einer Theorie₁ allgemeine Aussagen über das Verhältnis von zwei oder mehr Variablen („a general proposition, or a logically-connected system of general propositions“, Abend, 2008, S. 177). Theorien₁ seien in diesem Bedeutungszusammenhang aufgrund ihrer allgemeinen Gültigkeit vergleichsweise abstrakt formuliert und folglich – anders als beispielsweise die Befunde vieler explorativ-qualitativer Forschungsarbeiten – nicht mehr kontextgebunden.

Theorie₂: In ähnlicher Weise bezeichneten Theorien₂ Aussagensysteme, die soziale Phänomene kausal zu erklären vermögen. Das Ziel bestehe darin, eine Beobachtung möglichst eindeutig auf ihre Ursache(n) zurückzuführen. Im Falle komplexer Theorien resultiere eine Theorie₂ logisch aus mehreren Theorien₁, die sich jeweils auf einen Ursache-Wirkungs-Zusammenhang beziehen. Die Nähe der ersten beiden Theoriebegriffe zum wissenschaftlichen Positivismus ist unschwer erkennbar, wenn Abend (2008) schließlich fordert: „This explanation [in the sense of theory₂] should identify a number of ‘factors’ or ‘conditions’, which individually should pass some sort of counterfactual test for causal relevance, and whose interaction effects should be somehow taken into account“ (S. 178).

Theorie₃: Im Gegensatz zu den ersten beiden Theoriebegriffen hätten Theorien₃ zum Ziel, Beobachtungen zu interpretieren und auf ihre Implikationen zu befragen. Theorien₃ fungierten dann als Deutung eines sozialen Phänomens und seien weniger quantifizierender, sondern stärker interpretativer Art: „What theories₃ offer is an original ‘interpretation’, ‘reading’ or ‘way of making sense’ of a certain slice of the empirical world” (Abend, 2008, S. 178).

Theorie₄: Als Beispiele für Theorien₄ benennt Abend (2008) „‘interpretations’, ‘analyses’, ‘critiques’, ‘hermeneutical reconstructions’, or ‘exegeses’“ (S. 179). Diese Textgattungen deuten bereits darauf hin, dass Theorien₄ vorwiegend in der literaturwissenschaftlichen Hermeneutik verortet sind. Folglich entstehen Theorien₄ im Modus der Reflexion und Rekonstruktion über einen Ausgangstext und beinhalten Aussagen über dessen Bedeutungsgehalt.

Theorie₅: Während sich die ersten drei Theoriebegriffe explizit auf empirische Daten stützen und Aussagen über Phänomene oder Objekte treffen, bieten Theorien₅ Deutungsrahmen für die Interpretation von Wirklichkeit im Sinne einer „Weltanschauung“ (Abend, 2008, S. 179; Hervorhebung im Original): „[T]heories₅ focus on our conceptual and linguistic equipment – for example, the nature of the location from which we look at the social world, the lexicon and syntax by means of which we talk about it [...]“ (Abend, 2008, S. 179). Zudem zeichneten sich Theorien₅ dadurch aus, dass der Wahrheitsgehalt der in ihnen gesammelten Aussagen häufig nicht überprüfbar sei. Stattdessen fungierten Theorien₅ als Rahmenkonzepte oder „‘theoretical approach’“ (Abend, 2008, S. 180), welche die Wahrnehmung der Wirklichkeit – und damit folglich auch die Empirie – vorstrukturierten.

Theorie₆: Theorien₆ umfassten all solche theoretischen Ansätze, die eine stark normative Setzung beinhalten. Im Gegensatz zu den bisher abgegrenzten Theoriebegriffen, die einen eher deskriptiven Charakter aufwiesen, seien Theorien₆ präskriptiv. Deskriptive Ansätze versuchten, sich auf die Beschreibung von Fakten zu beschränken (z. B. „Der Schüler spricht das Wort ‘thunder’ aus wie /ˈsʌn.dər/.“); präskriptive Ansätze beinhalteten hingegen implizit oder explizit wertende Anteile und trafen Aussagen darüber, wie etwas sein sollte (z. B. „Der Schüler muss an seiner Aussprache von ‚th‘ arbeiten.“). In dem Beispiel erfolgt die Handlungsempfehlung aufgrund einer Wertung der deskriptiven Aussage entlang der *Native-Speaker*-Norm. Dass Abend (2008) für Theorien₆ – die in gewisser Hinsicht eine Teilmenge von Theorien₅ bilden – einen eigenen Theoriebegriff definiert, lässt Rückschlüsse auf seine kritische Haltung zu Theorien₆ zu: „[T]he word ‘theory’ can be used to refer to a normative, and indeed political, account – a far cry from other senses of it“ (S. 180).

Theorie₇: Schließlich umfassten Theorien₇ all solche Vorhaben, die eine Gegenstandsbetrachtung auf einer Meta-Ebene vornähmen und die Abend (2008) als „‘philosophical’ problems“ (S. 181) bezeichnet: „Even though it is because of its being in the business of empirically investigating society that sociology has encountered these problems, they are not empirical problems themselves (for example, they cannot be resolved by means of empirical methods)“ (S. 181). Theorien₇ erschlossen neue Perspek-

tiven auf einen Gegenstand, indem sie Interpretationen oder Analysen dessen (Theorien₄) einander gegenüberstellten oder auf neue Weise zusammenführten.

3.3 Teleologische Dimension: Wozu dient Theorie?

Obwohl die Zwecke von Theorien in den von Abend (2008) identifizierten Theoriebegriffen nicht von zentraler Bedeutung sind, lassen sich auf den semantischen Ebenen Indizien für primäre Funktionen des jeweiligen Theoriebegriffs finden. Wenn beispielsweise Theorien₂ dadurch charakterisiert sind, dass sie Aussagen über Ursache-Wirkungs-Zusammenhänge treffen, dann tragen sie zur Generierung von Erklärungswissen bei. In Antwort auf die Frage nach dem Zweck von Theorien unterscheiden Sandberg und Alvesson (2021) fünf Arten von Theorien anhand ihrer primären Funktion für den Erkenntnisgewinn (*explaining, comprehending, ordering, enacting, provoking*), die wir zur Klärung der teleologischen Dimension heranziehen. Auch hier betonen die Autoren, Theorien ließen sich nicht trennscharf einer der fünf Funktionen zuordnen. Da die Autoren in der Organisationsforschung arbeiten, verweisen sie in ihrem Aufsatz auf Theorien aus diesem Feld. Aufgrund der Verortung dieses Feldes in den Sozialwissenschaften gehen wir jedoch davon aus, dass die fünf Theoriefunktionen auf die Fremdsprachendidaktik übertragbar sind, was in Abschnitt 4 zu zeigen sein wird. Zuvor wird die Typologie in stark reduzierter Form wiedergegeben.

Explaining: Trage eine Theorie dazu bei, Wissen über Kausalzusammenhänge zu generieren, mit dem das Entstehen oder Wirken von Phänomenen beleuchtet werden könne, dann erfülle sie die Funktion des Erklärens:

„It [the theory] needs to demonstrate what variables make up the phenomenon, and how and why they are related. Most critically, explaining theory should, as far as possible, reflect and explain reality, preferably with causal relations clearly stated. Lastly, it needs to be empirically testable, typically in the form of rigorously developed and verifiably hypotheses“ (Sandberg & Alvesson, 2021, S. 497; Anmerkung der Verfasser).

Erklärende Theorien unterlägen stets der Annahme, dass die Phänomene, die sie zum Gegenstand haben, durch das Zusammenspiel von Variablen determiniert seien. Infolgedessen lassen sich erklärende Theorien eindeutig einem positivistisch geprägten Theorieverständnis zuordnen und dürften am häufigsten unter Theorien₁₋₂ auszuzeichnen sein.

Comprehending: Theorien dieser Funktion verfolgten das Ziel, zum genaueren Verständnis eines beobachteten Phänomens beizutragen: „The main *purpose* of comprehending theory, then, is to offer a qualified understanding of organizational [and other] phenomena by determining their meaning: that is, what phenomena [...] are about“ (Sandberg & Alvesson, 2021, S. 499; Hervorhebung im Original). Es scheint kein Zufall zu sein, dass die Autoren zur Definition dieses Theorietyps dieselbe Formulierung verwenden, wie es Abend (2008) in seinen Ausführungen zu Theorien₃ tut („Rather, you may mean that your theory helps understand what [...] ‘is all about’“, S. 179). Dies gelinge Theorien, indem sie auf sprachliche Strukturen wie Erzählungen, Diskurse oder Metaphern zurückgriffen. Theorien, die beim Verstehen eines Phänomens helfen, seien ge-

wöhnlich qualitativ-interpretativer Art und würden vor allem in hermeneutischer und phänomenologischer Forschung artikuliert. Dabei gelte die Konstruktion des Sinngehalts, der stets als ambig und vielschichtig verstanden wird, als konstitutives Moment des betrachteten Phänomens.

Ordering: Inwiefern solche Theorien positivistischer oder interpretativer Natur seien, hinge von den Verfahren ab, die zu ihrer Generierung eingesetzt würden. Denkbar seien dabei sowohl qualitative Verfahren wie die komparative Fallanalyse als auch quantitative Verfahren wie die explorative Faktorenanalyse. Gemeinsam sei all diesen Verfahren, dass sie die Komplexität betrachteter Phänomene zur Strukturierung reduzieren, ohne jedoch zentrale Merkmale aus den Augen zu verlieren (Sandberg & Alvesson, 2021, S. 500–501).

Enacting: Theorien dieses Typs setzten sich damit auseinander, wie Phänomene sozial konstruiert werden. Diesem Ziel näherten sie sich, indem sie versuchten „to articulate how phenomena are continuously produced and reproduced: that is, the processes through which they emerge, evolve, reoccur, change and decline over time“ (Sandberg & Alvesson, 2021, S. 502). Als Beispiel für diese Gruppe von Theorien sei auf das Konzept des *doing gender* (West & Zimmermann, 1987) verwiesen.

Provoking: Provozierende Theorien zeigten „alternative, often eye-opening and disruptive ways“ (Sandberg & Alvesson, 2021, S. 504) und regten so dazu an, ein Phänomen aus einer neuen Perspektive zu betrachten. Dies gelinge, indem etablierte Denkweisen irritiert und Gegenpole zu breit akzeptierten Theorien angeboten würden. Betont werde in diesem Kontext erneut die Perspektivität und das Vokabular von Forscher:innen, aber auch deren Kreativität in der Genese der Theorie. Dadurch gewannen provozierende Theorien an Nähe zu Abends (2008) Theoriens. Die Qualität der Provokation bemesse sich daran, inwiefern es ihr gelinge, Leser:innen von sich zu überzeugen und dazu zu bemächtigen, etablierte Denkmuster abzulegen. Wenn Theorien provozieren, liefern sie aus unserer Sicht auch immer neue Deutungsangebote für die Disziplin, die oftmals mit einer gewissen Aufbruchstimmung verknüpft sein mögen, aber stets auch weiterer Aushandlung bedürfen, wie tragfähig das jeweilige neue Angebot tatsächlich ist.

4 Exemplarische Verortung fremdsprachendidaktischer Forschungsarbeiten in der Heuristik

In den Abschnitten 3.1 bis 3.3 wurde unsere Heuristik mit ihren drei Dimensionen zur Systematisierung des Theoriebegriffs (vgl. Abbildung 1) erläutert. Im nächsten Schritt prüfen wir die Heuristik auf ihre Anwendbarkeit in der Fremdsprachendidaktik. Dazu präsentieren wir verschiedene Forschungsarbeiten, deren Erkenntnisinteresse wir jeweils knapp zusammenfassen, um sie auf dieser Grundlage in unserer Systematik zu verorten. Bei der Recherche von Forschungsarbeiten wurde darauf geachtet, ein Beispiel für jede Theorie₁₋₇ (semantische Dimension) anzuführen, welches anschließend entlang der beiden anderen Dimensionen (ontologische/teleologische Dimension) klassifiziert wurde. Die Auswahl der Forschungsarbeiten entspricht damit notwendi-

gerweise einer Teilmenge möglicher Theoriebegriffe aus der Heuristik und unterliegt der Perspektivität der beiden Verfasser mit ihren ganz individuellen – in vielerlei Hinsicht aber auch einander komplementären – Forschungsschwerpunkten und den damit verbundenen blinden Flecken.

Theorie₁, explaining, hoher Formalisierungsgrad: Teimouri und Kolleg:innen (2019) berichten eine Metaanalyse des Zusammenhangs zwischen der *L2 anxiety* von Schüler:innen und ihrer Leistung im Fremdsprachenunterricht. Auf Grundlage eines systematischen Reviews sichten die Autor:innen 97 englischsprachige Forschungsarbeiten aus 23 Ländern. Zur Untersuchung des beschriebenen Zusammenhangs mitteln sie 216 Korrelationskoeffizienten aus 105 unabhängigen Stichproben. Die Autor:innen interpretieren die mittlere Effektstärke von $r = -0,36$ als Beleg für ihre Hypothese: „Anxiety, overall, is negatively associated with L2 achievement“ (Teimouri et al., 2019, S. 377).

Diese Interpretation betrachten wir als Theorie₁ mit erklärendem Charakter. Mit den beiden Konstrukten der *L2 anxiety* und der Leistungsfähigkeit im Fremdsprachenunterricht treffen die Autor:innen eine Aussage über das Verhältnis zweier Variablen zueinander, dessen Stärke und Gerichtetheit sie mit Hilfe empirischer Daten bestimmen. Die Theorie zur Wirksamkeit von *L2 anxiety* ist hoch formalisiert, da sie durch ein lineares statistisches Modell mit zwei latenten Variablen beschrieben wird. Das gewählte Verfahren der Metaanalyse erlaubt die Berücksichtigung von Stichproben aus unterschiedlichen Kontexten (z. B. hinsichtlich des Alters der Schüler:innen oder der getesteten Fremdsprache) und resultiert dadurch in einer vergleichsweise allgemeinen Aussage. Zudem trägt der metaanalytische Befund zu einem besseren Verständnis über die Relevanz von affektiven Schülermerkmalen für die Leistungsentwicklung im Fremdsprachenunterricht bei.

Theorie₂, explaining, hoher Formalisierungsgrad: Phänomene, die Theorien₂ kausal zu begründen versuchen, zeichnen sich durch einen hohen Grad an Komplexität aus, der in einzelnen Forschungsarbeiten notwendigerweise nicht allumfassend betrachtet werden kann. Aus diesem Grund beschränken wir uns in der Illustration dieses Theoriebegriffs mit dem *age of onset* nicht auf eine einzelne Studie, sondern auf einen zentralen Gegenstand der Forschung zum frühen Fremdsprachenlernen, der inzwischen wiederholt bearbeitet wurde (Jaekel et al., 2017, 2022; Baumert et al., 2020; Wilden et al., 2020; Porsch et al., 2023).

Wir betrachten die oben angeführten Studien als Beispiele für erklärende Theorien₂. Obwohl die Arbeiten auf unterschiedlichen Stichproben und Instrumenten basieren und jeweils eigene Schwerpunkte in der Untersuchung des Forschungsgegenstandes setzen, beziehen sie sich mit dem Angebots-Nutzungs-Modell schulischer Lehrlernszenarien auf einen gemeinsamen theoretischen Bezugspunkt. Für das Fremdsprachenlernen folgt aus diesem Modell (das selbst wiederum auf unterschiedlichen *theoretischen* Annahmen im Sinne von Theorien₁ basiert), dass ein möglichst früh einsetzender Fremdsprachenunterricht die bloße Menge formaler Lerngelegenheiten erhöht und sich so positiv auf die fremdsprachliche Kompetenz der Schüler:innen auswirkt. Diese modellbasierte Annahme überprüfen die Referenzarbeiten hoch formali-

siert anhand längsschnittlicher Analysen. So ermöglichen sie einen kumulativen Erkenntnisgewinn darüber, inwiefern früheinsetzender Fremdsprachenunterricht seine theoretisch formulierten Vorteile einzulösen vermag.

Theorie₃, comprehending/ordering, niedriger Formalisierungsgrad: Wilken (2021) untersucht in ihrer rekonstruktiven Forschungsarbeit, welche handlungsleitenden Wissensbestände den Umgang von Englischlehrkräften mit Mehrsprachigkeit strukturieren. Dazu analysiert sie einen Korpus aus 16 episodischen Interviews unter Anwendung der Dokumentarischen Methode. Über alle Teilnehmer:innen hinweg rekonstruiert die Autorin den Umgang mit Korrektheit im Englischunterricht als handlungsleitende Habitusdimension: „Korrektheit und die damit verwobenen institutionellen Normen [...] konkurrieren mit der Einbeziehung von Mehrsprachigkeit und setzen sich durch, so dass die Norm der Korrektheit dem systematischen Einbezug von Mehrsprachigkeit und innovativen Unterrichtskonzepten im Wege steht“ (Wilken, 2021, S. 168).

Wilkens (2021) Hypothesen zu handlungsleitenden Wissensbeständen von Englischlehrkräften können als Theorie₃ typisiert werden, da sie eine in der praxeologischen Wissenssoziologie verankerte Interpretation dafür anbieten, „warum eine positive Einstellung zur Einbeziehung von Mehrsprachigkeit nicht handlungsleitend wird“ (S. 168). Die Forschungsarbeit leistet so einen Beitrag zu dem Verständnis, wie Englischlehrkräfte Mehrsprachigkeit im Spannungsfeld von Habitus und institutionellen Normen begegnen. Darüber hinaus resultiert die komparative Fallanalyse in einer sinn-gene-tischen (S. 150) und relationalen Typologie (S. 153), die Wilkens (2021) Kernbefunde strukturiert ordnen und für Anschlussprojekte zugänglich machen. Die Theoriearbeit hier ist gering formalisiert, da sie mit sozial konstruierten Gegenständen operiert, explorativ angelegt ist und folglich stark kulturell bedingt ist.

Theorie₄, comprehending/provoking, niedriger Formalisierungsgrad: Mihan (2018) präsentiert eine literaturdidaktische Arbeit, in der die unterrichtliche Zieldimension einer „critical gender literacy“ (S. 209) durch die Aushandlung von Transgender-Perspektiven mittels Williamsons Roman *The Art of Being Normal* erreicht werden soll. Zum einen analysiert Mihan (2018) in einem Modus des *close readings* zentrale Episoden des Romans in Bezug auf die dort auffindbaren Repräsentationen von Gender, insbesondere auch von „trans issues“ (S. 215), wodurch das besondere Potenzial dieses literarischen Werks für den Unterricht präzise herausgearbeitet wird. Zum anderen entwickelt Mihan (2018) einen Legitimationsrahmen zur Inklusion von Transperspektiven in den Englischunterricht, indem sie literaturdidaktische Forschungsdiskurse zu Gender sowie curriculare Dokumente kritisch überprüft und für die dort lokalisierbare Relevanz und Anschlussfähigkeit von *trans issues* argumentiert. Die Ergebnisse dieser hermeneutischen Arbeitsweise münden in die Artikulation einer in der Praxis erprobten Unterrichtskonzeption.

Diese Arbeit stellt insofern eine Theorie₄ dar, als Mihan unter Wahrung der Ambiguität des Ausgangstexts hermeneutisch begründet ein Interpretationsangebot desselben rekonstruiert. Dieses Angebot ist gering formalisiert, da es narrativ dargestellt und damit sprachlich bedingt ist. Darüber hinaus entwickelt Mihan, ebenfalls in hermeneutischer Arbeitsweise mit geringem Formalisierungsgrad, eine tiefgehende Re-

flexion dazu, wie Genderperspektiven bereits im Forschungsdiskurs ausgehandelt werden, wobei die hier entwickelte Kritik affin zu Theorie₄ erscheint. Die durch die *Gender Studies* und Transperspektiven informierte literaturwissenschaftliche und -didaktische Analyse trägt zu einem vertieften und differenzierten Verständnis von *trans issues* und genderbezogenen Lernzielen im Sinne des *comprehending* bei. Schließlich enthalten Mihans Interpretation und Unterrichtskonzeption auch ein provozierendes Element, indem sie Perspektiven auf Gender eröffnen, die sich Heteronormativität, Binarität und Cis-Geschlechtlichkeit deutlich widersetzen – und damit auch den Genderdiskurs in der Fremdsprachendidaktik zu irritieren vermögen.

Theorie₅, enacting/provoking, variierender Formalisierungsgrad: In ihrer Dissertation setzt sich Tödter (2023) mit dem Begriff der Fremdheit auseinander, der als konstitutives Element der Fremdsprachendidaktik in der Forschung oft implizit resoniert, aber selten explizit adressiert wird. Über die theoretisch angelegte Klärung der Frage, was fachimmanent und interdisziplinär unter *Fremdheit* verstanden werden kann, erreicht Tödter eine nuancierte Betrachtung, wie sich die Erfahrung von Fremdheit im Unterricht ergebnisoffen und komplex inszenieren lässt. Damit erweitert sie den Forschungsdiskurs, der bisher vor allem *Fremdheitsrepräsentationen* fokussiert hat, um den konkreten Blick auf unterrichtliche *Fremdheitserfahrungen*. Als Ergebnis legt Tödter ein didaktisches Modell mit handlungsleitenden Prinzipien zur Inszenierung von Fremdheitserfahrungen im Englischunterricht vor. Dieses Modell ist das Ergebnis einer hermeneutisch angelegten Erweiterung des Fremdbegriffs sowie einer in kleinerem Umfang durchgeführten qualitativen Befragung von Lehrpersonen als Expert:innen der unterrichtlichen Inszenierung von Fremdheit.

Aus unserer Sicht erscheint eine Zuordnung zu Theorie₅ adäquat, da Tödter sich über den Begriff der Fremdheit mit einem zentralen konzeptuellen *equipment* der Disziplin auseinandersetzt und dadurch einen Diskursrahmen schafft, der eine Aushandlung dieses Begriffs ermöglicht. Gleichzeitig vermag die Studie den Fremdbegriff vorzustrukturieren, um ihn empirisch weiter zu untersuchen. Tödters teleologische Theoriedimension entspricht dem *enacting*, da sie mit Fremdheit ein zentrales Phänomen der Fremdsprachendidaktik in Richtung von *Fremdheitserfahrungen* transformiert. Durch diesen transformatorischen Gehalt ist Tödter gleichzeitig auch *provoking*, da sie etablierte Denkweisen produktiv irritiert und ein neues Deutungsangebot zur unterrichtspraktischen Umsetzung derselben liefert. Schlussendlich ist der Formalisierungsgrad dieses Referenzbeispiels variierend: Zum einen ist die diskursiv-konzeptionelle Nuancierung des Fremdbegriffs typisch für einen geringen Formalisierungsgrad, während das empirisch abgeleitete Prinzipien-Modell einen höheren Grad an Formalisierung aufweist, aber weder hypothesenprüfende Funktion hat noch wissenschaftlich gedacht ist.

Theorie₆, enacting/provoking, niedriger Formalisierungsgrad: Braselmann (2024) widmet sich der Auswahl und Implementierung literarischer Texte im Englischunterricht aus einer rassismuskritischen Perspektive. In ihrer Argumentation arbeitet sie heraus, wie die Auseinandersetzung mit *Black-American-Young-Adult-Romanen* die Selbstreflexivität von (angehenden) Lehrkräften im Sinne einer *racial literacy* fördern kann. Als

Teil dessen entwickelt die Autorin Reflexionsfragen zu unterrichtsbezogenen Haltungen und Praktiken von Fremdsprachenlehrkräften sowie zur Auswahl und Implementation literarischer Texte.

Wir ordnen Braselmanns (2024) Aufsatz als Theorie₆ ein, welche die Konstruktion und (Re-)Produktion von *Whiteness* im Fremdsprachenunterricht Englisch nachzeichnet (*enacting*) und diesen Diskurs durch das Aufzeigen einer neuen Perspektive am Beispiel eines literarischen Genres zu irritieren versucht (*provoking*). Die Referenzarbeit nimmt im Rahmen dessen normative Setzungen vor, indem beispielsweise „the role of self-reflexivity and an awareness of the teacher’s positionality and racial identity as a prerequisite for teaching“ (Braselmann, 2024, S. 61) bezeichnet wird. Erst durch eine Reflexion der eigenen Unterrichtspraxis und die darin implizierten Überzeugungen gelinge es Fremdsprachenlehrkräften, *Whiteness* als dominanten, defizitären und unterdrückenden Diskurs zu erkennen und dem entgegenzuwirken. Des Weiteren wird gleich an mehreren Stellen auf die sozial konstruierte und diskursive Natur der Forschungsgegenstände hingewiesen: „[R]ace is understood as a social construct and a product of social thought [...]“ (Braselmann, 2024, S. 59; siehe auch S. 56, 66). Charakteristisch für einen niedrigen Formalisierungsgrad dieser Theorie ist damit primär der Soziolekt als Teil ihrer sprachlichen Bedingtheit. Die rassismuskritische Analyse trägt dazu bei, den Wert von *Black-American-Young-Adult*-Romanen in der Dezentrierung von *Whiteness* zu erkennen. Schließlich dokumentiert sich das Potenzial zur Provokation bereits auf sprachlicher Ebene in der Figur der „counter-stories“ (Braselmann, 2024, S. 66). Die Referenzarbeit bildet demnach ein konkretes Beispiel für die seitens der kritischen Fremdsprachendidaktik formulierte Theorie₆, „dass Sprache Macht ist und Sprache machtvoll machen kann, dass Sprache Ungleichheit konstruieren, diese aber auch relativieren kann, dass Sprache diskriminieren kann, aber auch davon erlösen kann“ (Gerlach, 2020, S. 7).

Theorie₇, *enacting/provoking*, *niedriger Formalisierungsgrad*: Als Beispiel für Theorie₇ führen wir Schmenks (2022b) kritische Reflexion des *Native-Speaker*-Konstrukts an. Ausgehend von der Streichung des *Native Speaker* im *Gemeinsamen Europäischen Referenzrahmen für Sprachen* (GER) arbeitet die Autorin Schwierigkeiten in der Konstruktdefinition heraus und fasst den Forschungsdiskurs zusammen, bevor sie die Entstehungsbedingungen des Konstrukts reflektiert und dessen Implikationen für Bildungspolitik und den Fremdsprachenunterricht formuliert.

Repräsentativ für diese Theorieklasse ist an der Analyse, dass sie mehrere Diskurse zum *Native-Speaker*-Konstrukt metaanalytisch zusammenfasst und interpretiert (die ihrerseits wiederum Theorien₄ darstellen), um es aus unterschiedlichen Perspektiven zu beleuchten. Die Arbeit erfüllt die Funktion des *enacting*, da sie durch die Systematisierung des Forschungsdiskurses die kulturhistorischen Entstehungsbedingungen sowie die institutionelle Wirkmacht des Konstrukts nachzeichnet. Zudem wird das provozierende Moment der Arbeit insbesondere an den Stellen deutlich, an denen die Autorin die Implikationen des Konstrukts für den Fremdsprachenunterricht einordnet. Schmenks (2022b) konzeptionelle Reflexion ist diskursiver Art und weist einen niedrigen Formalisierungsgrad auf, da sie ihren Ausgangspunkt in konkreten Fällen (z. B.

Streichung des *Native Speaker* im GER, *Native Speaker* in Ausschreibungen) nimmt, eine sorgfältige sprachliche Analyse der Forschungsliteratur umfasst und nicht zuletzt eine klare Einordnung des *Native-Speaker*-Konstrukts vornimmt.

5 Potenziale und Herausforderungen einer differenzierten Verwendung des Theoriebegriffs

Das zentrale Anliegen des vorliegenden Aufsatzes bestand darin, die Unschärfen des Theoriebegriffs in der Fremdsprachendidaktik zum Gegenstand forschungsmethodologischer Diskussion zu machen. Dazu wurde eine Heuristik entwickelt, die drei Dimensionen des Theoriebegriffs ausdifferenziert und anhand fremdsprachendidaktischer Forschungsarbeiten illustriert. Abschließend werden nun die Heuristik kritisch reflektiert, ihre Potenziale und Herausforderungen geprüft und Desiderata für die Theoriearbeit in der Fremdsprachendidaktik abgeleitet.

Inhalt und Struktur der entwickelten Heuristik zum Theoriebegriff werden maßgeblich von den Bezugstheorien (Abend, 2008; Zima, 2017; Sandberg & Alvesson, 2021) bestimmt. Dadurch ist grundsätzlich sichergestellt, dass die Heuristik Bezug auf Systematisierungsansätze nimmt, die sich in anderen Disziplinen bereits bewährt haben. Gleichzeitig liegt darin ein limitierendes Moment, indem andere Überlegungen zum Theoriebegriff ausgeschlossen werden. Zur Identifikation möglicher Bezugstheorien wurde eine breit angelegte Literaturrecherche zum Theoriebegriff durchgeführt. Die Mehrheit der identifizierten Publikationen setzte sich mit der Frage auseinander, was eine wissenschaftliche Theorie auszeichnet (z. B. Alexander, 1982; Thomas, 1997; Levine, 2017). Diese Frage wurde in Abschnitt 3.1 unter der ontologischen Theoriedimension adressiert. Alle Arbeiten zu dieser Fragestellung zu synthetisieren hätte eines weiteren Aufsatzes bedurft, weshalb die Frage danach, was Theorie sei, in der vorgelegten Heuristik auf den Formalisierungsgrad einer Theorie reduziert wurde. Die beiden wissenschaftstheoretischen Zugänge zum Theoriebegriff – Theorie als streng formalisiertes, falsifizierbares Aussagensystem (Popper, 1934) und Theorie als Dialog (Zima, 2017) – bilden die Pole eines Theoriekontinuums und erlauben so, die Vielzahl an Arbeiten zur Beschaffenheit von Theorien abzubilden. Die Fokussierung auf das Merkmal der Formalisierung ist insofern gerechtfertigt, als Theoriearbeit in der Fremdsprachendidaktik häufig diskursiv erfolgt und Zimas (2017) Entwurf einer Dialogischen Theorie den Wert dieser Herangehensweise dezidiert herausarbeitet.

In der exemplarischen Analyse fremdsprachendidaktischer Arbeiten zeigt sich aber auch, dass der Formalisierungsgrad sprachlich-diskursiver Theorien variiert. Braselmanns (2024) Theoriearbeit nimmt narrative Form an, wohingegen Wilken (2021) ihre Kernbefunde in einer mehrdimensionalen sinngenetischen Typenbildung systematisiert und abstrahiert. Da die gewählten Arbeiten die Vielfalt sprachlicher Darstellungsformen von Theorie aufgrund ihrer Beispielhaftigkeit nicht abbilden können, bestünde eine Aufgabe fremdsprachendidaktischer Forschung zukünftig darin, qualitative Abstufungen der ontologischen Dimension des Theoriebegriffs zu identifizieren.

Angesichts dessen sei erneut der tentative Charakter und die Nicht-Abgeschlossenheit der Heuristik betont. Es ist nicht ausgeschlossen, dass sich die drei Dimensionen der Heuristik durch weitere Klassifikationsmerkmale ergänzen lassen.

Des Weiteren ist darauf hinzuweisen, dass sich die Heuristik ausschließlich dazu eignet, den Inhalt des Theoriebegriffs präzise zu bestimmen. Sie kann von Forscher:innen bemüht werden, um zu schildern, wovon sie sprechen, wenn sie sich des Theoriebegriffs bedienen. Damit ist jedoch noch keine Aussage darüber getroffen, (1) durch welche Qualitätskriterien sich der jeweilige Theoriebegriff auszeichnet und (2) was es auf operativer Ebene bedeutet, Theoriearbeit in eben diesem Sinne zu betreiben. Aufgrund der Vielfalt der in Abschnitt 4 illustrierten Theoriebegriffe erscheint es legitim anzunehmen, dass sich diese beiden Fragen nicht allgemein beantworten lassen, sondern separat für jede Kombination aus Dimensionen des Theoriebegriffs zu diskutieren sind. Während es beispielsweise für theoretisch-konzeptionelle Arbeiten (z. B. Theorien₄) an forschungsmethodologischen Handreichungen mangelt, liegen für empirische Arbeiten (z. B. Theorien₁₋₃) zahlreiche Anleitungen zur Datenerhebung und -auswertung vor (z. B. Caspari et al., 2022).

Über diese Herausforderungen hinaus birgt die Heuristik das Potenzial, in der Fremdsprachendidaktik Klarheit im Sprechen *über* Theorie zu schaffen und die in Forschungsarbeiten vorgenommene Theoriearbeit zu explizieren. Dies sei beispielhaft an dem in der Fremdsprachendidaktik durchaus prominenten Diskurs zum Verhältnis von Theorie und Empirie (Matz & Rumlich, 2023; Römhild et al., 2024) demonstriert. Findigen Leser:innen ist sicher nicht entgangen, dass die Trennung von theoretisch-konzeptioneller und empirischer Forschung auch in diesem Beitrag an einigen Stellen implizit auftaucht, obgleich versucht wurde, Theoriearbeit nicht auf theoretisch-konzeptionelle Forschung zu begrenzen. In der Tat ist *Theorie* nicht mit *theoretisch-konzeptioneller Forschung* gleichzusetzen und ebenso wenig ist anzunehmen, dass nur theoretisch-konzeptionelle Arbeiten (*nomen est omen?*) mit Theorien operieren oder diese entwickeln. Zugleich sieht sich dieses Forschungsparadigma im Zuge der empirischen Wende zunehmend mit der Herausforderung konfrontiert, *ex negativo* als nicht-empirische Forschung definiert zu werden (Bonnet, 2017; Bellmann, 2020). Somit kommt es bisweilen zu gegenseitigen Vorwürfen von „theorieloser Empirie“ und „empirieloser Theorie“ (Vogd, 2005, S. 112). Unsere Heuristik vermag diesen Dualismus aufzulösen, indem sie explizit macht, dass diesen Vorwürfen unterschiedliche Theoriebegriffe zugrunde liegen. Wer Theoriearbeit einen Mangel an Empirie unterstellt, setzt implizit ein szientifisches Theorieverständnis zur Norm. Eine solche Annahme ist jedoch nur im Kontext von Theorien₁₋₃ gültig; denn Theorien₄₋₇ erheben gemäß der Heuristik erst gar nicht Anspruch darauf, dem Gütekriterium der Falsifizierbarkeit standzuhalten. Theorien₁₋₃ heben das Gegensatzpaar von Theorie und Empirie sogar völlig auf, da Theorie hier zu einem Produkt einer methodisch kontrollierten Auseinandersetzung mit Daten wird. Diskussionen über das Verhältnis von Theorie und Empirie sind folglich nur dann zielführend, wenn alle Forscher:innen auf dasselbe Theorieverständnis rekurrieren. Dies wird mit Hilfe der oben formulierten Heuristik erleichtert.

In Abschnitt 4 wurde veranschaulicht, dass sich ausgewählte fremdsprachendidaktische Forschungsarbeiten in der Heuristik zum Theoriebegriff verorten lassen. Daher sollte zukünftig kritisch geprüft werden, inwiefern sich die Vielfalt fremdsprachendidaktischer Forschung auf die drei Dimensionen der Theorieheuristik reduzieren lässt und inwiefern einzelne Forschungsarbeiten, aber auch größere Theorieströmungen, durchaus mehreren Kategorien innerhalb dieser Dimensionen zuzuordnen wären und Theoriebegriffe und -funktionen somit nicht disjunkt angelegt sind. Es wäre beispielsweise vorstellbar, Arbeiten zu einem bestimmten Forschungsgegenstand einer qualitativen Inhaltsanalyse zu unterziehen, um herauszuarbeiten, welche Merkmalskombinationen besonders häufig oder selten vorkommen. Außerdem ist aus unserer Sicht eine Diskussion darüber interessant, ob manche Teildisziplinen der Fremdsprachendidaktik affiner zu bestimmten Theoriebegriffen sein werden als andere und welche Implikationen sich daraus für die Weiterentwicklung der Theoriearbeit in unterschiedlichen Teildisziplinen ableiten lassen. Auf diese Weise ließe sich prüfen, inwiefern die Heuristik für Forscher:innen von praktischem Nutzen ist.

Die Frage nach dem praktischen Nutzen stellt sich überdies insbesondere in einer anwendungsbezogenen Wissenschaft wie der Fremdsprachendidaktik. Die vorliegende Analyse des Theoriebegriffs ist wissenschaftstheoretisch geprägt und mit dem Ziel verfasst worden, einen Beitrag zum forschungsmethodologischen Diskurs der Fremdsprachendidaktik zu leisten. Während eine differenzierte Auseinandersetzung mit dem Theoriebegriff keine konkreten Empfehlungen für die Gestaltung von Fremdsprachenunterricht liefert, wurden für den Kontext der Fremdsprachenlehrkräftebildung bereits erste Potenziale identifiziert (für Beiträge zu Neurowissenschaften und Spracherwerb, Mehrsprachigkeit und (vor-)theoretischem Wissen im Fremdsprachenunterricht, siehe Harant et al., 2020). Die Schwerpunktsetzung des Beitrags soll weder die Gegenüberstellung von Theorie und Praxis verfestigen noch den Wert von Theorien *für* die sowie *aus* der Praxis schmälern. Ein Charakteristikum wissenschaftlicher Theorien besteht jedoch gerade in der präzisen sprachlichen Explikation (Legutke & Scharf, 2025), die unsere Heuristik ermöglichen könnte.

Der Theoriebegriff bezeichnet ein unscharfes gedankliches Konstrukt, das sich bisweilen einer präzisen Definition entzieht. Wie versucht wurde darzustellen, ist die Präzisierung des Theoriebegriffs mehrdimensional und nicht immer eindeutig. Umso wichtiger erscheint es, zukünftig miteinander auszuhandeln, was ein tragfähiger Theoriebegriff in der Fremdsprachendidaktik, insbesondere mit Blick auf disziplinäre Besonderheiten, leisten muss, denn: „Theorie lebt vom Interesse für das Andere und Andersartige; sie erstarrt im Monolog“ (Zima, 2017, S. 68). Daraus ergibt sich die Notwendigkeit für Dialoge im Forschungsdiskurs, in denen keine Theoriepositionierung als absolut gedacht wird, sondern stets ein Deutungsangebot für die Disziplin darstellt. Ebenso kann Teil eines reflektierenden Dialogs sein, wie sich Theorien bestimmter Typen – auch reziprok zu sich wandelnden Gesellschaften – dynamisch entwickeln. Außerdem schlagen wir vor, dass zukünftige Forschungsarbeiten explizit(er) darlegen, wie sie jeweils in den Dialog mit *Theorie* treten. Mit unserer Heuristik des Theoriebegriffs möchten wir zum Gelingen dieses facettenreichen Dialogs beitragen.

Literaturverzeichnis

- Abend, G. (2008). The meaning of 'Theory'. *Sociological Theory*, 26(2), 173–199. <https://doi.org/10.1111/j.1467-9558.2008.00324.x>
- Alexander, J. (1982). *Positivism, presuppositions, and current controversies. Theoretical logic in sociology* (Band 1). University of California Press.
- Baumert, J., Fleckenstein, J., Leucht, M., Köller, O., & Möller, J. (2020). The long-term proficiency of early, middle, and late starters learning English as a foreign language at school: A narrative review and empirical study. *Language Learning*, 70(4), 1091–1135. <https://doi.org/10.1111/lang.12414>
- Bender, M., & Müller, M. (2020). Heuristische Textpraktiken in den Wissenschaften. *Zeitschrift für Germanistische Linguistik*, 48(1), 1–46. <https://doi.org/10.1515/zgl-2020-0001>
- Bellmann, J. (2020). Theoretische Forschung. *Zeitschrift für Pädagogik*, 66(6), 788–806. <https://doi.org/10.3262/ZP2006788>
- Bonnet, A. (2017). Forschungsmethoden und Forschungsinstrumente. In C. Surkamp (Hrsg.), *Metzler Lexikon Fremdsprachendidaktik* (S. 81–86). J. B. Metzler.
- Braselmann, S. (2024). A self-reflexive and critical approach to de-centering whiteness in teacher education. *Anglistik*, 35(2), 55–79. <https://doi.org/10.33675/ANGL/2024/2/8>
- Caspari, D. (2022). Grundfragen fremdsprachendidaktischer Forschung. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Hrsg.), *Forschungsmethoden in der Fremdsprachendidaktik* (S. 7–21). Narr. <https://doi.org/10.24053/9783823394327>
- Caspari, D., Klippel, F., Legutke, K. M., & Schramm, K. (Hrsg.) (2022). *Forschungsmethoden in der Fremdsprachendidaktik*. Narr. <https://doi.org/10.24053/9783823394327>
- Döring, N. (2023). Operationalisierung. In N. Döring (Hrsg.), *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (S. 223–291). Springer. https://doi.org/10.1007/978-3-662-64762-2_8
- Gadenne, V. (2017). Karl Poppers kritischer Rationalismus heute. In G. Franco (Hrsg.), *Handbuch Karl Popper* (S. 751–770). Springer. https://doi.org/10.1007/978-3-658-16239-9_50
- Gerlach, D. (Hrsg.) (2020). *Kritische Fremdsprachendidaktik*. Narr.
- Harant, M. & Thomas, P. (2020). Theorie – was sie war, wozu sie wurde und was sie heute in der Lehrerbildung sein kann: Historische und systematische Perspektiven. In M. Harant, P. Thomas, & U. Küchler (Hrsg.), *Theorien! Horizonte für die Lehrerinnen- und Lehrerbildung* (S. 23–26). Universitätsbibliothek Tübingen. <https://doi.org/10.35468/hblb2020-070>
- Harant, M., Thomas, P., & Küchler, U. (Hrsg.) (2020). *Theorien! Horizonte für die Lehrerinnen- und Lehrerbildung*. Universitätsbibliothek Tübingen. <https://doi.org/10.35468/hblb2020-070>
- Jaekel, N., Schurig, M., Floria, M., & Ritter, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, 67(3), 631–664. <https://doi.org/10.1111/lang.12242>

- Jaekel, N., Schurig, M., & Ritter, M. (2022). Investigating cohort effects of early foreign language learning. *The Language Learning Journal*, 52(1), 104–116. <https://doi.org/10.1080/09571736.2022.2108124>
- Keuth, H. (2017). Karl Poppers „Logik der Forschung“. In G. Franco (Hrsg.), *Handbuch Karl Popper* (S. 45–63). Springer. https://doi.org/10.1007/978-3-658-16239-9_3
- Kleining, G. (2020). Qualitative Heuristik. Entdeckende Strategie mit qualitativen Daten in der Psychologie. In G. Mey, & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie*. (S. 203–223). Springer. https://doi.org/10.1007/978-3-658-18234-2_14
- König, L., Schädlich, B., & Surkamp, C. (2022). *unterricht_kultur_theorie: Kulturelles Lernen im Fremdsprachenunterricht gemeinsam anders denken*. J. B. Metzler. <https://doi.org/10.1007/978-3-662-63782-1>
- Legutke, M. K. (2022). Theoretische Forschung. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Hrsg.), *Forschungsmethoden in der Fremdsprachendidaktik* (S. 40–50). Narr.
- Legutke, M. K., & Schart, M. (2025). Theorie. In C. Surkamp (Hrsg.), *Metzler Lexikon Fremdsprachendidaktik* (S. 472–474). J. B. Metzler.
- Levine, D. (2017). *Social theory as a vocation*. Routledge. <https://doi.org/10.4324/9781315129822>
- Matz, F., & Rumlich, D. (16.03.2023). *Theorie vs. Empirie – eine forschungsmethodische Grundsatzdiskussion* [Konferenzbeitrag]. DGFF Nachwuchstagung 2023: Findungsphasen, Münster, Deutschland.
- Mihan, A. (2018). The Art of Being Normal – Trans issues in the EFL classroom. In M. Eisenmann, & C. Ludwig (Hrsg.), *Queer Beats – Gender and Literature in the EFL Classroom* (S. 207–229). Peter Lang.
- Popper, K. (1934). *Logik der Forschung*. Julius Springer.
- Porsch, R., Schipolowski, S., Rjosk, C., & Sachse, K. (2023). Effects of an early start in learning English as a foreign language on reading and listening comprehension in year 9. *Language Teaching for Young Learners*, 5(2), 122–148. <https://doi.org/10.1075/ltyl.22017.por>
- Reimann, D. (2020). *Methoden der Fremdsprachenforschung*. Narr.
- Römhild, R., Matz, F., & Merse, T. (05.07.2024). *Theoretische Forschung in der Fremdsprachendidaktik* [Symposium]. Münster, Deutschland.
- Sandberg, J., & Alvesson, M. (2021). Meanings of theory: Clarifying theory through typification. *Journal of Management Studies*, 58(2), 487–516. <https://doi.org/10.1111/joms.12587>
- Schmenk, B. (2022a). Textzusammenstellung (für theoretische Arbeiten). In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Hrsg.), *Forschungsmethoden in der Fremdsprachendidaktik* (S. 138–147). Narr.
- Schmenk, B. (2022b). Das *Native-Speaker*-Konstrukt im Fokus: Kritik und Kontroversen. *Zeitschrift für Fremdsprachenforschung*, 33(2), 163–187.

- Schmenk, B., Breidbach, S., & Küster, L. (2019). Sloganization in language education discourse: Introduction. In B. Schmenk, S. Breidbach, & L. Küster (Hrsg.), *Sloganization in Language Education Discourse. Conceptual Thinking in the Age of Academic Marketization* (S. 1–18). Multilingual Matters. <https://doi.org/10.2307/jj.22730494.4>
- Schramm, K. (2021). Forschungsansätze zur Didaktik/Methodik des Deutschen als Fremd- und Zweitsprache. In C. Altmayer, K. Biebighäuser, S. Haberzettl, & A. Heine (Hrsg.), *Handbuch Deutsch als Fremd- und Zweitsprache* (S. 215–232). J. B. Metzler. https://doi.org/10.1007/978-3-476-04858-5_14
- Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41, 363–387. <https://doi.org/10.1017/S0272263118000311>
- Thomas, G. (1997). What's the use of theory? *Harvard Educational Review*, 67(1), 75–104. <https://doi.org/10.17763/haer.67.1.1x807532771w5u48>
- Tödter, M. (2023). *Fremdheit im Englischunterricht. Eine interdisziplinäre Grundlagenarbeit zur Gestaltung von Fremdheitserfahrungen*. WVT.
- VanPatten, B., Williams, J., Keating, G. D., & Wulff, S. (2020). Introduction: The nature of theories. In B. VanPatten, G. D. Keating, & S. Wulff (Hrsg.), *Theories in second language acquisition* (S. 1–18.). Routledge. <https://doi.org/10.4324/9780429503986-1>
- Vogd, W. (2005). Komplexe Erziehungswissenschaft jenseits von empirieloser Theorie und theorieloser Empirie: Versuch einer Brücke zwischen Systemtheorie und rekonstruktiver Sozialforschung. *Zeitschrift für Erziehungswissenschaft*, 8(1), 112–133. <https://doi.org/10.1007/s11618-005-0126-1>
- Volkman, L. (2022). Hermeneutische Verfahren. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Hrsg.), *Forschungsmethoden in der Fremdsprachendidaktik* (S. 240–252). Narr.
- West, C., & Zimmermann, D. H. (1987). Doing gender. *Gender and Society*, 1(2), 125–151. <https://doi.org/10.1177/0891243287001002002>
- Wilden, E., Porsch, R., & Schurig, M. (2020). An early start in primary EFL education and the role of teacher qualification and teaching quality. *Language Teaching for Young Learners*, 2(1), 28–51. <https://doi.org/10.1075/ltyl.19002.wil>
- Wilken, A. A. (2021). *Professionalisierung durch Schüler*innen-Mehrsprachigkeit? Englischlehrer*innen im Spannungsfeld zwischen Habitus und Norm*. Julius Klinkhardt.
- Zima, P. V. (2017). Was ist Theorie? Theoriebegriff und Dialogische Theorie in den Kultur- und Sozialwissenschaften. A. Francke. <https://doi.org/10.36198/9783838547978>

Diagnosing Cognitive Engagement in TEFL

Validating and Analyzing a Simulation

MERAL ROEBEN¹, CHRISTIANE LÜTGE², FLORIAN SCHULTZ-PERNICE³, JOHANNA VEJVODA⁴, MICHAEL SAILER⁵, FRANK FISCHER⁶ & NICOLE HEITZMANN⁷

Abstract

Kognitive Aktivierung zählt zu den drei Basisdimensionen guten Unterrichts. In dieser Studie konzeptualisieren wir sie mit dem ICAP-Modell, das vier Stufen unterscheidet, die durch Beobachtung von Schüler:innen-Aktivitäten bestimmt werden können. Wir wenden das Modell auf den Englischunterricht an und nutzen eine Simulation, in der Lehramtsstudierende und Lehrkräfte im Dienst lernen, die Stufen kognitiver Aktivierung beim Planen und Realisieren von Unterricht zu diagnostizieren. Ziel war die Validierung der Simulation sowie die Replikation der fachübergreifenden Studie von Roeben et al. (2025), um die kontextübergreifende Anwendbarkeit des Modells zu prüfen. Mit N = 118 Lehramtsstudierenden einer bayerischen Universität fanden wir Belege für die Validität. Zudem zeigt die Signalentdeckungstheorie, dass sich die Übertragung in den englischdidaktischen Kontext unter Einschränkungen als erfolgversprechend erweist. Wir empfehlen daher, die Doppelstruktur des Fremdsprachenlernens in das Modell aufzunehmen und es um typische Beispiele zu erweitern.

-
- 1 Meral Roeben, Research associate, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, Department of English and American Studies, Chair of Teaching English as a Foreign Language, Schellingstraße 3, 80799 Munich, Germany, E-Mail: Meral.Roeben@lmu.de
 - 2 Christiane Lütge, Professor, LMU Munich, Department of English and American Studies, Chair of Teaching English as a Foreign Language, Schellingstraße 3, 80799 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Christiane.Luetge@lmu.de
 - 3 Florian Schultz-Pernice, Academic director, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Florian.Schultz-Pernice@psy.lmu.de
 - 4 Johanna Vejvoda, Research associate, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Johanna.Vejvoda@psy.lmu.de
 - 5 Michael Sailer, Professor, University of Augsburg, Faculty of Philosophy and Social Sciences, Department of Learning Analytics and Educational Data Mining, Alter Postweg 101, 86135 Augsburg, Germany, E-Mail: Michael.Sailer@uni-a.de
 - 6 Frank Fischer, Professor, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Frank.Fischer@psy.lmu.de
 - 7 Nicole Heitzmann, Assistant Professor, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Nicole.Heitzmann@psy.lmu.de

1 Introduction

One of the three basic dimensions of good teaching is cognitive engagement (Praetorius & Gräsel, 2021). This study investigates diagnosing cognitive engagement in teaching English as a foreign language (TEFL). Diagnostic skills are part of teachers' professional knowledge (Kramer et al., 2021). In modern TEFL classrooms, diagnosing is often conducted in the context of technology-related teaching. This process is complex and presupposes respective diagnostic skills. Simulations have shown to effectively support acquiring complex skills by reducing the complexity of a diagnostic situation and focusing only on aspects relevant for diagnosing (Heitzmann et al., 2019). In the present study, we aim at validating a simulation that supports acquiring diagnostic skills regarding cognitive engagement and instructional TEFL quality in technology-related TEFL lessons. Additionally, we aim at empirically investigating the concept of cognitive engagement within the context of foreign language teaching. More specifically, we investigate challenges that occur during diagnosing cognitive engagement in the TEFL context by replicating parts of a study conducted in a cross-domain context (Roeben et al., 2025). We hope to derive theoretical and practical implications for TEFL by comparing the patterns found in both studies.

1.1 Diagnostic Skills

Teachers make a variety of complex decisions which should be informed by aspects like the students' knowledge, motivation, or emotions (Kramer et al., 2021; Urhahne & Wijnia, 2021). Diagnosing such phenomena involves identifying a problem and finding a solution to it (Heitzmann et al., 2019). Based on these diagnoses, teachers can design content and tasks that fit students' needs, thus support their learning process (Urhahne & Wijnia, 2021). The respective diagnostic skills consist of conceptual knowledge (e. g., knowing the characteristics of a framework) and action-oriented knowledge (e. g., knowing how to put a framework to practice; Kopp et al., 2008). Today, teaching increasingly involves the use of technology. Technology within this article refers to any computer-based technology that supports teaching and learning. When making diagnostic decisions, teachers need to assess the potential of technology in enhancing students' learning. Technology, however, can also be used to inform the diagnostic processes, as technology allows insights into students' current activities from which phenomena central to the learning process, such as cognitive engagement, can be inferred.

1.2 Types of Knowledge Required in Different Phases of Teaching

Teachers diagnose cross-domain and subject-specific phenomena in technology-related teaching situations. The professional knowledge required to do so, is conceptualized by the TPACK model (Koehler et al., 2013). The TPACK model describes seven domains of knowledge: Content knowledge (CK) refers to knowledge about a subject's content, while pedagogical knowledge (PK) includes effective teaching methods; integrated, these types of knowledge form pedagogical content knowledge (PCK; Koehler et al., 2013). Technological knowledge (TK) refers to knowledge about how to use and

implement technology and the integration of CK, PK, and TK results in technological pedagogical and content knowledge (TPACK; Koehler et al., 2013). It is the teacher's task to combine the different domains and base the teaching on this combined knowledge (Koehler et al., 2013).

To provide students with beneficial learning opportunities, teachers apply TPACK in the different phases of teaching (Ertmer & Ottenbreit-Leftwich, 2010). In the present paper, we focus on diagnosing in the planning phase (i. e., planning lessons) and in the implementation phase (i. e., giving lessons in the classroom). We assume that the respective diagnostic decisions differ: In the planning phase, learning goals and tasks are diagnosed regarding aspects like cognitive engagement. Diagnosing in the implementation phase includes making decisions within the dynamic system of a classroom. In the present study, during the planning phase, participants diagnose phenomena related to PCK (see 1.4; Müller-Hartmann, 2017) and PK (see 1.3; Chi & Wylie, 2014), in the implementation phase a phenomenon related to PK (see 1.3; Chi & Wylie, 2014) is diagnosed. Results regarding the respective diagnostic performance (i. e., accurately diagnosing the different phenomena) in combination with conceptual knowledge inform about the validity of the simulation designed in the course of this study (see RQ1 in chapter 2).

1.3 Cognitive Engagement

One component of PK is being able to diagnose cognitive engagement – one of the core dimensions of good teaching – which is featured in prominent instructional quality models, such as the “Syntheseframework” and the MAIN-TEACH model (Praetorius & Gräsel, 2021). Cognitive engagement alludes to students' depth of cognitive involvement with instructional material (Praetorius & Gräsel, 2021). One promising way of conceptualizing this is the ICAP framework which distinguishes four observable levels of cognitive engagement – interactive, constructive, active, and passive – and assumes a hierarchy in terms of cognitive processing depth (Chi & Wylie, 2014). In contrast, the construct of cognitive activation (Klieme & Rakoczy, 2008) refers to internal cognitive stimulation through instructional design, which is not directly observable. This distinction is particularly relevant in TEFL, where communication plays a central role in learning and presupposes deep involvement with the language (Wilden, 2021). However, within the TEFL context the conceptualization of cognitive engagement varies greatly (Guttke, 2023). We acknowledge that the ICAP framework, while empirically established and offering diagnostically accessible levels of cognitive engagement, does not fully capture these deeper and subject-specific facets of cognitive activation. However, given the difficulty of conceptualizing cognitive engagement in cross-disciplinary contexts, our study adopts ICAP as a workable framework for empirical investigation, while making its assumptions and limitations transparent. By investigating the ICAP framework in the TEFL context, we aim to provide a nuanced understanding of cognitive engagement in TEFL teaching practice.

The ICAP framework assumes that each level of cognitive engagement describes a certain observable student activity (Chi et al., 2018): The passive level of cognitive en-

gement refers to activities in which students do not get visibly active (e.g., reading, watching). However, the passive level does not imply the absence of learning – aligning with the constructivist notion that learning is never fully passive. The passive level merely refers to the absence of any other activity than attending to information presented by the teacher. In the ICAP framework, this level of activity is associated with low level cognitive process of storing information (Chi & Wylie, 2014). The active level is characterized by students being physically active but not generating new knowledge beyond the instructional material (e.g., copying what the teacher has written at the blackboard). The constructive level includes generating new knowledge, beyond the presented learning material (e.g., creating a mind-map based on the content of a video). The interactive level is reached when knowledge is co-generated (i.e., due to the exchange of ideas between two people, new knowledge emerges). For instance, a mind-map could be created jointly. The interactive level is only reached if the new knowledge is a product of the exchange. Thus, if students merely generate knowledge on their own and share it but do not build on the knowledge of the other students, they have not reached the interactive level.

Chi and Wylie (2014) note that although the ICAP framework assumes that the levels can be determined based on observation, distinguishing the active, constructive, and interactive level additionally requires inferring from student products. Consequently, we assume that these levels are more challenging to differentiate than to distinguish between passive and the other levels. (Roeben et al., 2025). This aligns with Chi et al. (2018) reporting that teachers struggled with differentiating active and constructive levels and with designing truly interactive learning activities.

However, there is little research on applying the ICAP framework in foreign language teaching. The present study examines how the context influences the difficulty of diagnosing levels of cognitive engagement, focusing on both the planning and implementation phase of technology-related TEFL lessons.

1.4 Instructional Quality of Tasks in Teaching English as a Foreign Language

One aim of this study is to validate the simulation we designed which is called Digivate-E. In order to do so, we assessed diagnostic skills regarding the instructional quality of tasks in TEFL lesson plans.

In contrast to cross-domain teaching criteria, evidence for subject-specific – especially TEFL-specific – criteria is limited (Praetorius & Gräsel, 2021; Wilden, 2021). In the present study, we conceptualize criteria for instructional quality of tasks in TEFL lesson plans with the core principles of modern TEFL proposed by Müller-Hartmann (2017) who bases his principles – action-orientation, interculturality, learner-centeredness, task-orientation, meaningful content, and self-regulated and cooperative learning – on the Common European Framework of Reference (Council of Europe, 2001). As participants were assumed to vary in their prior knowledge, the relevant information on these aspects of TEFL was integrated as supplementary material within the simulation and could be accessed any time.

1.5 Simulation-based Learning

Diagnosing cognitive engagement and instructional quality of technology-related TEFL lessons, both when planning and implementing them, requires diagnostic skills. As these are complex skills, they may be trained as early as in the first phase of teacher education. Simulations facilitate the acquisition of complex skills and offer the opportunity of imitating and approximating teaching practice to a certain degree while adapting the complexity to learners' current skill level (Grossman et al., 2009). Typical characteristics of simulations include the reduction in complexity of the presented situation and the opportunity for learners to interact with the simulation (Heitzmann et al., 2019). In order to ascertain that a simulation enables learners to acquire the intended knowledge and skills, it needs to be validated. Common criteria for validation are that individuals who dispose of pertinent conceptual knowledge perform better than those who do not (Kane, 2006). An additional indicator is intrinsic cognitive load – assuming that learners with lower pertinent knowledge will experience higher intrinsic cognitive load as the tasks are more difficult for them (Klepsch et al., 2017).

2 The Present Study⁸

The first aim of this study is to validate the simulation Digivate-E to ensure that it is effective for learning and to provide a valid basis for further research questions. We hope to gain insights into the predictiveness and correlations of conceptual PK, PCK, CK, TEFL-specific professional knowledge (i. e., PCK and CK; Kirchhoff, 2017) and the respective action-oriented PK and PCK (i. e., diagnostic skills in the respective domains). Moreover, by replicating parts of the study by Roeben et al., 2025, our second aim is to provide systematic research on the difficulty of diagnosing the different levels of cognitive engagement in a field different from the originally investigated one. These insights could point us towards necessary scaffolding or cues which are needed in the simulations to make it adaptive and thus more effective for learners with different learning prerequisites (Plass & Pawar, 2020).

Consequently, we pose the following research questions:

- RQ1:** To what extent can evidence for the validity of the simulation be found? To address this question, we investigate in detail:
- RQ1.1:** To what extent can we reliably assess action-oriented pedagogical knowledge and pedagogical content knowledge within the simulation Digivate-E?

⁸ The first author's contribution by Meral Roeben was funded by a grant from the Hanns Seidel Foundation. This research was further funded by the European Union (Next Generation EU)—and by the German Federal Ministry of Education and Research under the grant number 01JA23S01E.

- RQ1.2:** To what extent is conceptual knowledge (i. e., TEFL-specific professional, pedagogical knowledge) predictive for action-oriented knowledge (i. e., pedagogical content, pedagogical knowledge) as assessed within the simulation?
- RQ1.3:** To what extent is the intrinsic cognitive load predictive for action-oriented pedagogical knowledge within the simulation?
- H1.1:** We hypothesize that action-oriented pedagogical and pedagogical content knowledge show acceptable reliability within the simulation.
- H1.2a:** We hypothesize that conceptual TEFL-specific professional knowledge is predictive for action-oriented pedagogical content knowledge and action-oriented pedagogical knowledge.
- H1.2b:** We hypothesize that conceptual pedagogical knowledge is predictive for action-oriented pedagogical knowledge.
- H1.3:** We hypothesize that a lower reported intrinsic cognitive load is predictive for higher action-oriented pedagogical knowledge within the simulation.
- RQ2:** To what extent does the difficulty of diagnosing levels of cognitive engagement depend on differences in the levels of cognitive engagement (inferring vs. no inferring) within the phases of teaching (i. e., planning phase, implementation phase)?
- H2a:** We hypothesize that levels of cognitive engagement with no need of inferring (passive) are easier to distinguish than levels that require inferring (active, constructive, interactive).
- H2b:** We hypothesize that the difficulty of diagnosing the levels of cognitive engagement is different in the planning and the implementation phase.

3 Methodology

3.1 Sample and Design

The present study is an observational study with a correlational design. Between June 2023 and December 2023 we collected data sets of $N = 162$ pre-service teachers (all school types), studying English at a Bavarian university (degree: state exam), of which we could use $N = 118$ complete data sets (no missing data). Of the participants, 71 % were female, 27 % male, and 2 % diverse. On average, they had taught 37 lessons ($M = 37.20$; $SD = 86.89$). Of the three TEFL modules, participants have to take during their studies, about half of the study participants had completed the first module and half the second one ($M = 1.54$, $SD = .71$). The study participants were recruited through advertisement in seminars and on university websites. The study was conducted as a laboratory study; the participants could earn 36€ for participating. The laptops the participants conducted the study on, were provided and set up by the researchers.

3.2 Learning Environment and Participants’ Tasks

For the present study, we adapted the previous simulation Digivate (Roeben et al., 2025) and developed a simulation in the context of TEFL – Digivate-E. Digivate-E can be accessed via a website and is built as a point-and-click-adventure using comic-style visuals, audios, videos, and text documents. Study participants took on the role of teacher trainees conducting their teacher training at a secondary school in a class in their third year of learning English as a second language. The class is currently reading the graded Klett teamreader *The Magic Mirror* by Josh Lacey (2019).

In their role of teacher trainees, the study participants are greeted by the seminar teacher who introduces them to their first set of tasks (planning phase; see Fig. 1). This first task consists of looking at existing lesson plans on the sequence of *The Magic Mirror*. The study participants are told to diagnose the potential level of cognitive engagement of the learning goals and of the tasks within lesson plans. They are also asked to determine the quality of the tasks from a TEFL perspective. After a total of twelve lesson plans are diagnosed, the seminar teacher introduces the participants to the second phase, the implementation phase (see Fig. 2). The seminar teacher asks the participants to accompany her to a classroom and observe the students working on the tasks. While observing them, the study participants determine the students’ current level of cognitive engagement. The student activity is represented by screen-videos of the students’ tablets or phones.

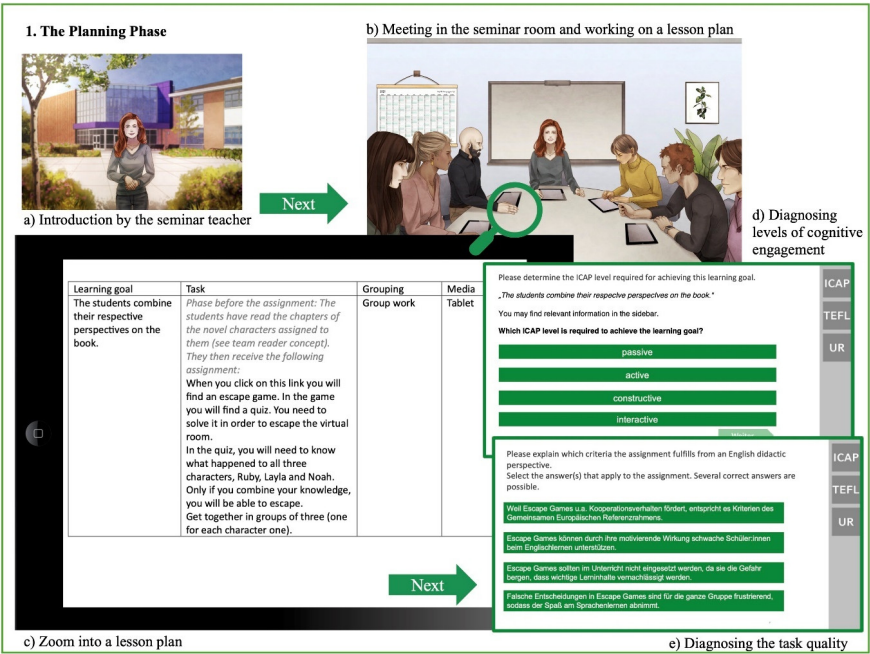


Figure 1: Overview of the Structure of the Planning Phase. (Source: “Overview of the Structure of the Planning Phase”, created by Meral Roeben, pictures drawn by Nina Ploch, licensed as CC BY SA 4.0)

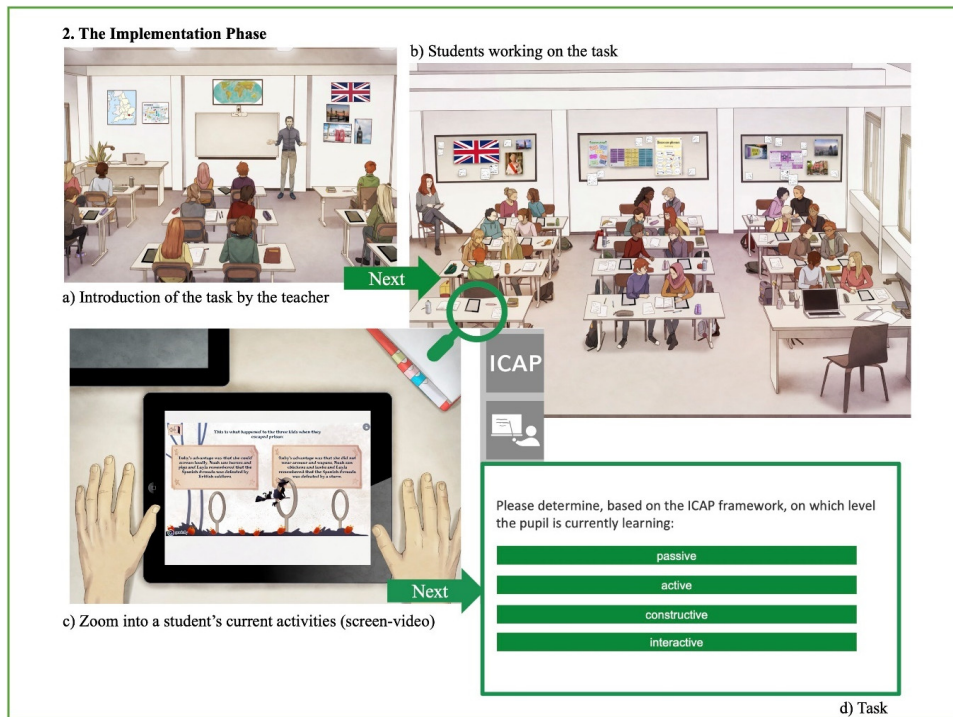


Figure 2: Overview of the Structure of the Implementation Phase (Source: “Overview of the Structure of the Implementation Phase”, created by Meral Roeben, pictures drawn by Nina Ploch, licensed as CC BY SA 4.0)

In the following, we speak of cases. One case in the planning phase consists of analyzing one lesson plan (i. e., evaluating the potential level of cognitive engagement of either the learning goal or the task and justifying the TEFL quality of the task), and one case in the implementing phase consists of diagnosing the current level of cognitive engagement of one student. The cases were presented in the same sequence to all participants.

Before starting on their first case, the study participants completed a comprehensive test on conceptual PCK and CK. During their first case of the planning phase, the study participants had access to a video explaining the ICAP framework. Participants had 25 minutes to complete their first case (i. e., test case) in the planning phase. Afterwards, the participants were tested on conceptual knowledge on the ICAP framework. After that, the participants worked on eleven more cases for which they were granted six minutes each to solve them. Subsequently, in the implementation phase, the participants again first worked on a test case (ten minutes) and on eleven further cases (six minutes per case). In both phases, the participants were asked to report their intrinsic cognitive load after they had determined students' level of cognitive engagement (i. e., one validation claim; see 1.5).

We assumed participants' prior knowledge to differ as the sample included pre-service teachers from different semesters and studying different school types. There-

fore, during the whole study, we offered additional information on the ICAP framework (slides from the video that was available during the first case) and on TEFL (see 1.4) in a sidebar. In the implementation phase, they also found information on the respective student they were diagnosing.

3.3 Measures

Intrinsic cognitive load was measured on a seven-point Likert scale, conceptual knowledge with open ended and closed questions, action-oriented knowledge as the accuracy of diagnostic decisions within the simulation, levels of cognitive engagement as levels of the ICAP framework, and the difficulty of diagnosing these levels by sensitivity and specificity.

3.3.1 Intrinsic Cognitive Load

The intrinsic cognitive load was measured in both phases. In the planning phase, it was measured after the level of cognitive engagement was determined for the learning goal or task and before the TEFL quality of the task was diagnosed. It was measured on a seven-point Likert scale (“do not agree at all” to “fully agree”) for the two questions: “When assessing this lesson plan/student activity, I had to work on many things in my head at the same time.” and “The assessment of this lesson plan/student activity was very complex.” (Klepsch et al., 2017).

3.3.2 Conceptual Knowledge (TEFL-specific professional knowledge, PCK, CK, PK)

Conceptual knowledge was assessed by different tests. Conceptual PCK and conceptual CK knowledge was assessed by the FALKO-E test (Kirchhoff, 2017). FALKO-E consists of two parts. One part is a test on PCK (Cronbach’s $\alpha = .50$) and one is a test on CK (Cronbach’s $\alpha = .55$), both for the TEFL context. When combining the two parts of the FALKO-E test, it assesses the conceptual TEFL-specific professional knowledge (Cronbach’s $\alpha = .68$; Kirchhoff, 2017). The PCK part of FALKO-E consist of 12 items, with two closed and ten open questions. For the CK part of the FALKO-E test, we assessed nine items, with four closed and five open questions. Conceptual knowledge on PK was assessed by eight single-choice items on the ICAP framework (Cronbach’s $\alpha = .68$).

3.3.3 Action-oriented Knowledge (PCK, PK)

To test PCK and PK in action, we assessed action-oriented PK and PCK within the simulation. We assessed the action-oriented PCK and PK as the participants’ performance, which was measured as the participants’ diagnostic accuracy. For action-oriented PCK, we measured the participants’ performance regarding diagnosing the TEFL-related instructional quality of the task during the planning phase of the simulation (“Please explain which criteria the task fulfills from an English didactic perspective. Select the answer(s) that apply/applies to the task. Several correct answers are possible.”). The participants could choose one or more answers from a total of four answer options per question. For each accurately selected and accurately not selected option, they gained a

point. Thus, per question, they could achieve four points and a total of 20 points in the planning phase.

To evaluate the action-oriented PK, we assessed the performance regarding accurately diagnosing the levels of cognitive engagement during the planning phase (“Please assess which ICAP level is required to achieve the following learning objective/task. You may find helpful information in the sidebar.”) and during the implementation phase (“Based on the ICAP framework, please assess the ICAP level the student is actually engaged on – regardless of which ICAP level the task was intended to stimulate!”). The participants could choose one of the four levels of cognitive engagement and gained one point for selecting the accurate level. They could reach eleven points in the planning phase as well as in the implementation phase, thus a maximum of 22 points for diagnosing the levels of cognitive engagement, as we did not take the first case of each phase into account for the performance measure.

3.3.4 Levels of Cognitive Engagement

Levels of cognitive engagement are conceptualized as the levels of the ICAP framework (i. e., passive, active, constructive, interactive). The accurate levels we assigned to each learning goal, task, and student activity were carefully validated in the complex process of an expert evaluation and workshop with researchers and practitioners. For diagnosing levels of cognitive engagement, we divided the levels. Active, constructive, and interactive levels require inferring from the student products to accurately determine them as merely observing students is not sufficient for these. In contrast, the passive level does not require inferring from the student product but can be determined by observing students because, in contrast to all other levels, students do not get physically active.

3.3.5 Difficulty of Diagnosing Levels of Cognitive Engagement

We assessed the difficulty of diagnosing the levels of cognitive engagement by creating confusion matrices which in the present study compare the accurate level (predicted) of cognitive engagement to the selected level (actual) of cognitive engagement. This way, we can find out which levels of cognitive engagement tend to get confused with each other. Based on confusion matrices, sensitivity and specificity can be calculated. Sensitivity is the likelihood that the accurate level is selected. If it is easy to diagnose a level, its sensitivity is high. Specificity is the likelihood that a level is not selected when it is inaccurate. If a level is easy to diagnose, its specificity is high.

3.3.6 Statistical Analyses

For RQ1, we conducted simple linear regressions and Pearson correlations to assess the associations between conceptual knowledge and action-oriented knowledge within the simulation. Additionally, we conducted simple linear regressions to check whether a reported lower intrinsic cognitive load is predictive for higher performance within the simulation (i. e., higher action-oriented PK).

To address RQ2, we made use of the signal detection theory which measures skills by considering the accurate and wrong answers (Wixted, 2020). This way, errors can be

understood better, and interventions or scaffolds can be used to prevent these errors from reoccurring (Wixted, 2020). So-called confusion matrices display how often a level of cognitive engagement was selected, for instance, if passive is the accurate level, how often was passive, active, constructive, or interactive selected. Based on this, the likelihood that an answer is accurately selected or accurately not selected can be calculated. In the present study we created confusion matrices for both phases of teaching (see Table 2): for each level of cognitive engagement in the planning phase (see Table 3 and 4), and for each level of cognitive engagement in the implementation phase (see Table 5). For the planning phase, we conducted a further differentiation by creating separate confusion matrices for learning goals (see Table 3) and tasks (see Table 4). One confusion matrix depicts the accurate level (predicted) on the x-axis and the selected level (actual) on the y-axis. Thus, the diagonal of the matrix displays the accurately selected accurate levels. These are called True Positives. Within one column, for example the active column, the incorrect levels (passive, constructive, interactive) are called False Negatives. To calculate the sensitivity, the True Positive value is divided by the sum of True Positive and False Negative values. To calculate the specificity, all occasions on which an incorrect level was accurately not selected (True Negative) are divided by the sum of the True Negative value and the False Positive value. False Positives describe all instances in which the level concerned was selected although a different level would have been the accurate choice.

Additionally, to explore whether certain levels of cognitive engagement were systematically confused more often than others, we conducted Chi-square tests, comparing the observed frequencies of confusion to the expected values under the assumption of independence. To assess whether there are significant differences between sensitivities and specificities for the different levels, we calculated confidence intervals using the Wilson method. There are significant differences between them if the confidence intervals of the sensitivity or specificity do not show any overlap.

4 Results

Regarding RQ1, validating the simulation, we summarized the descriptive results in Table 1. It shows that conceptual PCK is higher than conceptual CK, but conceptual PK exceeds all conceptual knowledge types. This is also true for the performance within the simulation as action-oriented PK is higher than action-oriented PCK.

Table 1: Standardized Descriptive Results of the Conceptual and Action-oriented Knowledge

Variables	N	M	SD	Min	Max
Conceptual PCK	118	.43	.14	.13	.75
Conceptual CK	118	.33	.14	.06	.67
TEFL-specific professional knowledge	118	.38	.12	.15	.71

(Continuing Table 1)

Variables	N	M	SD	Min	Max
Conceptual PK	118	.70	.25	.13	1.00
Action-oriented PCK	118	.68	.13	.33	.94
Action-oriented PK	118	.74	.11	.40	.95

Action-oriented PK showed an acceptable internal consistency (Cronbach's $\alpha = .66$) while it was low for action-oriented PCK (Cronbach's $\alpha = .43$; RQ1.1). Other validation claims are that a low intrinsic cognitive load is predictive for high performance (i.e., high level of action-oriented PK) and that conceptual knowledge is predictive for the performance within the simulation (i.e., higher action-oriented knowledge; RQ1.3). The intrinsic cognitive load was assessed after participants had diagnosed the level of cognitive engagement. Linear regressions show that the intrinsic cognitive load explains 8% of the variance in action-oriented PK ($R^2 = .08$) with a standardized regression coefficient for intrinsic cognitive load of $\beta = -.30$ ($p < .005$; H1.3). With regard to RQ1.2, we found that the conceptual PK explains 44% of the variance of the action-oriented PK ($R^2 = .44$) with a standardized regression coefficient for conceptual PK of $\beta = .67$ ($p < .001$; H1.2b). Moreover, conceptual TEFL-specific professional knowledge explains 8% of variance in action-oriented PK ($R^2 = .08$) with a standardized regression coefficient for conceptual TEFL-specific professional knowledge of $\beta = .29$ ($p < .001$; H1.2a). Regarding the action-oriented PCK, the conceptual TEFL-specific professional knowledge accounts for 11% in variance ($R^2 = .11$; $\beta = .34$, $p < .001$; H1.2a). The results of all linear regressions support our validity claim as they show moderate to strong, statistically significant associations of concept knowledge with action-oriented knowledge within the simulation. The residuals were approximately normally distributed, supporting the use of linear regression.

For the linear regressions, we combined conceptual PCK and conceptual PK in the construct of conceptual TEFL-specific professional knowledge. However, we also looked at these knowledge domains separately, conducting unidirectional Pearson correlations. Conceptual CK correlates higher with action-oriented PCK ($r = .39$, $p < .01$) than conceptual PCK correlates with action-oriented PCK ($r = .20$, $p < .05$). Another finding was that conceptual CK shows a significant correlation with action-oriented PK ($r = .36$, $p < .01$) while this is not the case for conceptual PCK ($r = .15$, n.s.). There is also no significant correlation between conceptual PK and action-oriented PCK ($r = .12$, n.s.). The assumption of normally distributed residuals was met.

To address RQ2, identifying the difficulty in diagnosing different levels of cognitive engagement in the planning and implementation phase of teaching, we created confusion matrices and calculated sensitivities and specificities. We created one confusion matrix including both the planning and implementation phase (see Table 2), two confusion matrices for the planning phase, namely for the learning goal (see Table 3) and the task (see Table 4), and one for implementation phase (see Table 5). We summed up the sensitivities and specificities overall (i.e., both phases), for learning goals

(i. e., part one of the planning phase), for tasks (i. e., part one of the planning phase), and for the student activities (i. e., the implementation phase) in Table 6.

While we did not include the test cases (i. e., first case of the planning and implementation phase) into the calculations regarding RQ1, we included them in the results of RQ2 as only the first case of the planning phase includes the passive level (i. e., the accurate level for the learning goal of the first case). Yet, due to this being the first case in the simulation, the respective results need to be treated with caution.

Overall (i. e., the planning and implementation phase; see Table 2) the True Positives (i. e., the occasions when the accurate level was correctly selected by participants) were the highest for all levels but for the passive level. Passive was most often diagnosed as active by participants. This confusion was confirmed as significant by a Chi-square test ($\chi^2(1) = 41.40$, $p < .001$). Active and constructive were frequently confused with each other. Chi-square tests confirm a significant association between the active and constructive level ($\chi^2(1) = 286.96$, $p < .001$). Interactive was sometimes confused with constructive, mostly however, it was accurately determined. This is reflected in the high sensitivity of interactive (see Table 6.). Overall, specificity was highest for passive and interactive (see Table 6). Confidence intervals for both sensitivity and specificity did not overlap, indicating that the sensitivities and specificities of all levels differed significantly from each other.

Table 2: Confusion Matrix for Diagnostic Accuracy in the Planning and Implementation Phase

		Predicted			
		P	A	C	I
Actual	P	39 (33 %)	55 (9 %)	15 (2 %)	4 (1 %)
	A	68 (58 %)	443 (75 %)	232 (28 %)	20 (3 %)
	C	5 (4 %)	87 (15 %)	439 (53 %)	60 (10 %)
	I	6 (5 %)	4 (1 %)	140 (17 %)	506 (86 %)

Note: The cells describe the overlap of predicted (i. e., accurate) levels and actual (i. e., selected) levels of cognitive engagement for all four levels (i. e., passive, active, constructive).

In the planning phase, we find similar patterns as we find overall (see Table 2, Table 3, Table 4) with frequent confusions of active and constructive and a high sensitivity for interactive.

In the planning phase, we will first look at the confusion matrix for learning goals (see Table 3). The True Positives (i. e., predicted level equals actual level) for passive and constructive are not the highest values. Both were most frequently determined as active. As mentioned before, for the passive level, this result is based on one case which was also the test case. As for constructive, this result is derived from four different cases, two rather in the beginning of the simulation (case 3, case 4) and two in the end (case 8, case 9). For active and interactive, the levels of cognitive engagement were most of the time determined accurately. While interactive was hardly ever confused, active was frequently

mistakenly classified as constructive. Chi-square tests confirm the significant association of active and constructive ($\chi^2(1) = 7.45, p < .05$). Sensitivity for learning goals is highest for interactive followed by active (see Table 6) with an overlap of confidence intervals for passive (.25–.42) and constructive (.30–.42), indicating that there is no significant difference in sensitivity between them. Specificity is highest for passive, followed by interactive with no overlap of any confidence intervals. In comparison to tasks (i. e., second part of the planning phase) sensitivity for learning goals is lower (see Table 6).

Table 3: Confusion Matrix for Diagnostic Accuracy for Learning Goals (Planning Phase)

		Predicted			
		P	A	C	I
Actual	P	39 (33 %)	2 (2 %)	6 (3 %)	0 (0 %)
	A	68 (58 %)	86 (73 %)	124 (53 %)	4 (2 %)
	C	5 (4 %)	28 (24 %)	84 (36 %)	30 (13 %)
	I	6 (5 %)	2 (2 %)	22 (10 %)	202 (86 %)

Tasks (see Table 4) are the second part in the planning phase. Here, the True Positives are the highest values for all levels. Again, the active and constructive levels are frequently confused which is confirmed by a Chi-square test ($\chi^2(1) = 145.39, p < .001$). Interactive is rarely confused, reflected in its high sensitivity (see Table 6). For sensitivity, there is no overlap of confidence intervals for any levels within tasks. Specificity is high for all levels with confidence intervals overlapping for active (.86–.91), constructive (.86–.91), and interactive (.89–.94), indicating that there is no significant difference between these specificities.

Table 4: Confusion Matrix for Diagnostic Accuracy for Tasks (Planning Phase)

		Predicted			
		P	A	C	I
Actual	P	NA	8 (3 %)	3 (1 %)	1 (0 %)
	A	NA	185 (79 %)	45 (19 %)	7 (3 %)
	C	NA	41 (17 %)	152 (64 %)	12 (5 %)
	I	NA	1 (0 %)	36 (15 %)	216 (92 %)

We will now focus on the implementation phase (i. e., the levels of cognitive engagements participants diagnosed for the student activities). In contrast to the planning phase, the pattern of confusing levels of cognitive engagement differs from the overall (i. e., planning and implementation phase) pattern. For the implementation phase, it is not the active and constructive levels that mostly get confused but instead the construc-

tive and interactive levels (see Table 5). Chi-square tests confirm a significant association between constructive and interactive ($\chi^2(1) = 90.33, p < .001$).

Table 5: Confusion Matrix for Diagnostic Accuracy for Student Activities (Implementation Phase)

		Predicted			
		P	A	C	I
Actual	P	NA	45 (19%)	6 (2%)	3 (3%)
	A	NA	172 (73%)	63 (18%)	9 (8%)
	C	NA	18 (8%)	203 (57%)	18 (15%)
	I	NA	1 (0%)	82 (23%)	88 (75%)

For sensitivity and specificity of interactive, we also find a new pattern in the implementation phase as both sensitivity and specificity are lower than overall and in the planning phase (see Table 6). This hints towards a higher difficulty in diagnosing interactive in the implementation phase. Confidence intervals for sensitivity overlap for active (.67–.78) and interactive (.66–.82). For specificity there is an overlap of confidence intervals between active (.81–.88) and interactive (.83–.89).

Table 6: Sensitivity and Specificity Overall, for Learning Goals, Tasks, and Student Activities

Overall			Planning Phase				Implementation Phase	
Both Phases			Learning Goal		Task		Student Activity	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
P	33 %	96 %	33 %	99 %	NA	98 %	NA	92 %
A	75 %	79 %	73 %	68 %	79 %	89 %	73 %	85 %
C	53 %	88 %	36 %	87 %	64%	89 %	57 %	90 %
I	86 %	90 %	86 %	94 %	92 %	92 %	75 %	86 %

5 Discussion

5.1 Summary of the Results

Regarding RQ1, we found that overall, the validity claim tested in the present study is supported: Conceptual knowledge (TEFL-specific professional knowledge, PK) is predictive for the respective action-oriented knowledge as shown in the performance in the simulation (PCK, PK; H1.2a; H1.2b). Yet, while conceptual PK explains a substantial proportion in variance of action-oriented PK, conceptual TEFL-specific professional knowledge shows a small effect size in predicting action-oriented PCK. Additionally, conceptual TEFL-specific professional knowledge is predictive for action-oriented PK

but shows a small effect size (H1.2a). Moreover, a low intrinsic cognitive load is predictive for higher performance regarding action-oriented PK, however the respective effect size is small (H1.3). We also found that the correlation between conceptual CK and action-oriented PCK is higher than the one between conceptual PCK and action-oriented PCK. Due to the low internal consistency of action-oriented PCK, both correlations may in fact even be stronger (Stadler et al., 2021; H1.1). Additionally, conceptual CK significantly correlates with action-oriented PK, while this is not the case for conceptual PCK.

For RQ2, we replicated Roeben et al. (2025) to find out whether the findings from the previous study also apply in a TEFL context. Overall, we found similar results: Over both phases and for the planning phase, confusion matrices show that the active and constructive levels of cognitive engagement are confused most frequently (H2). For the implementation phase, constructive and interactive levels are mostly confused (H2b). Moreover, in the implementation phase, the interactive level shows a comparatively low sensitivity (H2a). Passive, the level that does not need to be inferred from student products when determining it, shows high specificities overall and for both phases (H2a). However, in contrast to Roeben et al. (2025), we found that for learning goals, passive and constructive was surprisingly often mistakenly determined as active (i. e., True Positives < number of selected active levels). This matches the low sensitivity found for passive.

5.2 Practical and Theoretical Implications

Regarding RQ1, correlations suggest that the curriculum may place greater emphasis on CK-related courses and topics than PCK-related ones: There is a high correlation between conceptual CK and action-oriented PK as well as a significant correlation of conceptual CK and action-oriented PK. Consequently, it may cautiously be suggested that PCK is taught insufficiently at universities. However, our finding that the average conceptual CK is lower than the average conceptual PCK (see Table 1) challenges the assumption that the focus on PCK is insufficient. One explanation may be that the test on conceptual PCK is easier than the one on conceptual CK (Kirchhoff, 2017). PCK answers might be easier to transfer from general knowledge, while for CK a participant either has a particular piece of knowledge or cannot answer the question. From this, we may derive that a more reliable test on conceptual PCK knowledge is required. With a closer look at the questions on conceptual CK, we find that the questions on linguistics and literature are framed to be relevant to the school-context (Kirchhoff, 2017). This may indicate that teaching-related conceptual CK is more important for teaching than conceptual PCK.

Furthermore, the predictiveness of conceptual TEFL-specific professional knowledge for action-oriented PK indicates that in a TEFL context, cross-domain conceptual PK is not sufficient but additional TEFL-specific professional knowledge is required (Seidel & Shavelson, 2007) to diagnose cognitive engagement. Hence, a practical implication for the first phase of teacher education may be to integrate courses on PCK and PK more closely.

With RQ2, we replicated Roeben et al. (2025) and found similar results, indicating that the ICAP framework can be applied to the TEFL context. In contrast to the active level, the constructive level presupposes a knowledge generation process. Hence, this confusion (i. e., active with constructive) may indicate that participants struggled with grasping the concept of how or whether knowledge is generated (Chi et al., 2018). To address the challenges of identifying knowledge generation, scaffolds such as guiding questions that support pre-service teachers towards understanding this process could be implemented in an updated version of Digivate-E. Additionally, feedback on learners' solution may help reflect on the reasoning process. The confusion of the constructive and interactive level as well as the relatively low sensitivity of interactive may be due to an inaccurate overgeneralization. When observing an activity in which more than one student is involved, participants may jump to the conclusion that the students are engaged on an interactive level. However, it is possible that students have a conversation to which everyone contributes; nevertheless, not all or no student at all may be engaged on an interactive level because knowledge is generated on their own and not co-generated. Manipulating small aspects of the exchange between students that highlight whether the previous information is integrated in an answer, may help making the difference between a truly interactive interaction and one in which students remain on a constructive level more salient (Plass & Pawar, 2020). These adaptive and supporting elements may be based on metrics, such as the numbers of correctly or incorrectly solved cases, as well as on the time required to answer, or on detecting certain confusion patterns (i. e., confusing active and constructive or constructive and interactive). Both in this study and in Roeben et al. (2025) learning goals showed lower sensitivity in comparison to tasks. The levels within the ICAP framework are conceptualized in terms of student activities; hence, applying them to learning goals requires extensive inferential processes. Learning goals do not focus on a certain activity but describe which skill is supposed to be acquired with an activity. Inferring the level of cognitive engagement that can potentially be achieved by the activity that a learning goal aims at is more complex than inferring the levels for tasks or student activities. This complexity can be addressed by adding a meta-perspective to the framework, such as a step-by-step instruction to the diagnostic process. Moreover, Wekerle et al. (2024) suggest redesigning the levels, which are currently rather fixed categories, into fine-grained dimensions. Hence, the strict categories could be dissolved to a certain extent, allowing for more differentiation within the levels. These precise sub-dimensions could support understanding the activity better and facilitate diagnosing the respective learning goals as they may become more feasible. Both approaches may make the ICAP framework more comprehensive in terms of including not only the implementation but also the planning phase of teaching.

The present study showed new findings regarding diagnosing the levels from stated learning goals. For the passive and the constructive level, the majority of participants diagnosed the active level (i. e., active outnumbering the passive and constructive level). The confusion of the active and constructive level fits the overall confusion pattern. Yet, the confusion of passive and active level is novel and may hint towards certain

restrictions in applying the ICAP framework to the TEFL context. In contrast to subjects taught in students' mother tongue, in foreign language teaching the medium (i. e., the language) which is used to convey content is still being learned (König et al., 2016). Thus, language has two functions: It is both the content and the means of communication (Wilden, 2021). The levels of cognitive engagement within the ICAP framework describe observable student activities. Only when taking students' products into account may we be able to make more sophisticated statements about the students' covert cognitive processes. Yet, even then, the descriptions of the levels are merely focused on generating knowledge in terms of content. The medium is not considered, leading to the conclusion that the ICAP framework neglects mediality. This is especially problematic when it comes to novice foreign language learners. For them, coordinating those two functions may be challenging while later on, when the rules of language are more automatized, expert students have the capacity to focus mostly on the language as means of communication. Digivate-E, however, is set in the classroom of year seven when students are still learning the rules of a language. With that in mind, when determining the level of cognitive engagement of learning goals in the TEFL context, participants may analyze these learning goals in great depth and realize that there are two sides to a learning goal. For instance, the learning goal may consist in being able to read a certain text. This presupposes both, decoding skills (i. e., knowing the words) and linguistic comprehension (García & Cain, 2014). Participants may have struggled with considering such a complex goal as merely passive in the ICAP classification. In contrast, in their description of the ICAP framework, Chi and Wylie (2014) suggest reading as one typical example for a passive student activity. As just argued, when teaching a foreign language, reading is not an ideal example to describe a passive student activity, especially for novice students. Thus, one implication to fit the ICAP framework to the TEFL context may include considering the aspect of mediality, for instance by adding typical TEFL-specific activities to each level of the framework. At this point it may also be considered taking students' proficiency into account as reading an English text is more challenging for novice students and it is questionable whether novice learners are able to engage on a truly interactive level due to a lack of language skills. This dual perspective enables a more accurate interpretation of learners' engagement, especially when analyzing ambiguous activities such as reading, where cognitive demands may be underestimated by ICAP alone. To improve diagnostic accuracy, TEFL-specific indicators (e. g., task types, language level) should be considered in interpreting levels of cognitive engagement. This subject-specific refinement supports a more valid application of the ICAP framework in language teacher education.

5.3 Limitations

Although we validated the simulation and replicated results from a previous study, this study has limitations. Most importantly, due to its low internal consistency, interpreting results regarding action-oriented PCK should be treated with caution. The low consistency may stem from action-oriented PCK being a latent construct including various knowledge facets assessed within the simulation (Stadler et al., 2021). Additionally, the

criteria used to assess action-oriented PCK pose a limitation as empirically tested criteria are lacking. Reliable tests for (action-oriented) PCK are urgently needed.

The sample is another limitation, as the study was conducted with pre-service teachers from one Bavarian university. It would be interesting to test whether similar results emerge with pre-service teachers from other universities or with in-service teachers.

A further limitation concerns the number of passive cases. Assumptions toward the passive level are based on one case (i. e., the test case). The high confusion with the active level may thus be due to participants' unfamiliarity with the ICAP framework and simulation when working on the first case. Future studies may control for case sequence and include several cases for each level. Additionally, the findings for the passive level may imply that the ICAP framework in its current form is not able to capture the full complexity of cognitive engagement in the TEFL context. Thus, we suggest using an updated version that accounts for subject-specific particularities and better reflects the assumed depth structures of cognitive activation.

Finally, as Digivate-E was designed for validation, it does not allow participants to manipulate the simulation and thus lacks one characteristic of simulations (Heitzmann et al., 2019). A next step would be to test whether similar results occur in a version with more freedom to select or skip cases and choose additional scaffolding if needed.

5.4 Conclusion

Overall, the validity claims are supported, suggesting that Digivate-E is valid for the TEFL context and can be developed into an adaptive learning environment (Plass & Pawar, 2020). The study underscores the importance of subject-specific knowledge for cross-domain diagnostic skills (i. e., diagnosing cognitive engagement) and of integrating subject-specific and cross-domain pedagogical knowledge. Applying the ICAP framework – originally from cognitive psychology – to TEFL was largely successful: In both contexts, levels of cognitive engagement vary in diagnostic difficulty across teaching phases, and confusion patterns are similar. Still, revised versions of ICAP may need to address causes and remedies of diagnostic challenges, especially in TEFL, where adaptations are required. For instance, the framework should be extended by subject-specific activities, as it currently only lists cross-domain ones. The transfer to TEFL also enabled a more holistic analysis of language teaching, exposing the complexity of learning goals and their dual function: language as both content and medium (Wilden, 2021). While our data only allow us to infer this dual structure, it highlights that ICAP neglects mediality (i. e., language) more generally. This duality may also apply to other subjects, but further research is needed. Finally, the study provided methodological insights: By using signal detection theory (confusion matrices, sensitivity, specificity), we showed how it can be applied to evaluate teaching decisions.

References

- Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., Levy, R., Li, N., McElldoon, K. L., Stump, G. S., Wylie, R., Xu, D., & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science*, 42(6), 1777–1832. <https://doi.org/10.1111/cogs.12626>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. <https://rm.coe.int/1680459f97>
- Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change. *Journal of Research on Technology in Education*, 42(3), 255–284. <https://doi.org/10.1080/15391523.2010.10782551>
- García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, 84(1), 74–111. <https://doi.org/10.3102/0034654313499616>
- Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9), 2055–2100. <https://doi.org/10.1177/016146810911100905>
- Guttke, J. (2023). Kognitive Aktivierung im Fremdsprachenunterricht: Ein systematisches Review von Forschungsarbeiten aus dem deutschsprachigen Raum. *Zeitschrift für Fremdsprachenforschung*, 34(2), 145–176. <https://doi.org/10.17185/duerpublico/79249>
- Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. *Frontline Learning Research*, 7(4), 1–24. <https://doi.org/10.14786/flr.v7i4.384>
- Müller-Hartmann, A. (2017). Didaktische Leitideen, Konzepte und Prinzipien im Englischunterricht der Gegenwart. In F. Haß (Ed.), *Fachdidaktik Englisch* (2nd ed., pp. 26–30). Ernst Klett Sprachen.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). Praeger.
- Kirchhoff, P. (2017). FALKO-E: Fachspezifisches professionelles Wissen von Englischlehrkräften: Entwicklung und Validierung eines domänenspezifischen Testinstruments. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff, & R. H. Mulder (Eds.), *FALKO: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik* (pp. 113–152). Waxmann.
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01997>

- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222–237. <https://doi.org/10.25656/01:4348>
- Koehler, M. J., Mishra, P., & Cain, W. (2013). What is technological pedagogical content knowledge (TPACK)? *The Journal of Education*, 193(3), 13–19. <https://doi.org/10.1177/002205741319300303>
- König, J., Lammerding, S., & Nold, G. (2016). Teachers' professional knowledge for teaching English as a foreign language: Assessing the outcomes of teacher education. *Journal of Teacher Education*, 67(4), 320–337. <https://doi.org/10.1177/0022487116644956>
- Kopp, V., Stark, R., & Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: Effects of erroneous examples and feedback. *Medical Education*, 42(8), 823–829. <https://doi.org/10.1111/j.1365-2923.2008.03122.x>
- Kramer, M., Förtsch, C., Seidel, T., & Neuhaus, B. J. (2021). Comparing two constructs for describing and analyzing teachers' diagnostic processes. *Studies in Educational Evaluation*, 68, 100973. <https://doi.org/10.1016/j.stueduc.2020.100973>
- Lacey, J. (2019). *The Magic Mirror*. Ernst Klett Sprachen.
- Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, 52(3), 275–300. <https://doi.org/10.1080/15391523.2020.1719943>
- Praetorius, A.-K., & Gräsel, C. (2021). Noch immer auf der Suche nach dem heiligen Gral: Wie generisch oder fachspezifisch sind Dimensionen der Unterrichtsqualität? *Unterrichtswissenschaft*, 49(2), 167–188. <https://doi.org/10.1007/s42010-021-00119-6>
- Roeben, M., Vejvoda, J., Murböck, J., Fischer, F., Schultz-Pernice, F., Lohr, A., Stadler, M., Sailer, M., & Heitzmann, N. (2025). Simulations in Teacher Education: Learning to Diagnose Cognitive Engagement. *Education Sciences*, 15(3), 261. <https://doi.org/10.3390/educsci15030261>
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. <https://doi.org/10.3102/0034654307310317>
- Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, 60, 100832. <https://doi.org/10.1016/j.newideapsych.2020.100832>
- Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, 32, 100374. <https://doi.org/10.1016/j.edurev.2020.100374>
- Wekerle, C., Daumiller, M., Janke, S., Dickhäuser, O., Dresel, M., & Kollar, I. (2024). Putting ICAP to the test: How technology-enhanced learning activities are related to cognitive and affective-motivational learning outcomes in higher education. *Scientific Reports*, 14(1), 16295. <https://doi.org/10.1038/s41598-024-66069-y>
- Wilden, E. (2021). Fachspezifische Aspekte von Unterrichtsqualität im Schulfach Englisch. *Unterrichtswissenschaft*, 49(2), 211–219. <https://doi.org/10.1007/s42010-021-00105-y>

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 201–233. <https://doi.org/10.1037/xlm0000732>

Speaking Tests in the Lower Secondary Foreign Language Classroom

Students' Views on Usefulness, Anxiety, Preparation and Feedback

GABRIELA LÜTHI¹, ELISABETH PEYER² & NADIA RAVAZZINI³

Abstract

Diese Studie untersucht, welche Ansichten und Wünsche Schüler:innen der Sekundarstufe I bezüglich Sprechtests im Fremdsprachenunterricht haben. 254 Lernende aus der französisch- und italienischsprachigen Schweiz beantworteten schriftlich Fragen zu Vorbereitung, Feedback, Nützlichkeit und Stress während mündlicher Prüfungen im Fach Deutsch als Fremdsprache. Ebenfalls wurden Kurzinterviews mit 84 Lernenden direkt nach einer mündlichen Prüfung geführt und deren Lehrpersonen (N = 7) nach ihren Praktiken und Ansichten befragt. Rund zwei Drittel der Lernenden erachten verschiedene Prüfungsformate (Paarprüfungen, Einzelprüfungen, Vorträge) als nützlich, wobei sie dies vor allem mit der Wichtigkeit des Sprechens selbst begründen. Alle Prüfungsformate werden von einer Mehrheit jedoch auch als eher stressig empfunden, wobei die Paarprüfung sowohl als das am wenigsten stressige als auch als das nützlichste Format angesehen wird. Auffallend ist, dass die Lehrpersonen bezüglich der Nützlichkeit kritischer eingestellt sind als die Lernenden. Bezüglich Feedbacks bevorzugen die Schüler:innen individuelles Feedback, das ihnen zeigt, wie sie sich verbessern können. Allerdings ist das gegebene Feedback möglicherweise nicht direkt umsetzbar und es bleibt unklar, was genau die Lernenden damit machen (können).

1 Introduction

The assessment of speaking skills goes hand-in-hand with various challenges, including the fact that assessing an oral performance is complicated by the fleeting nature of spoken language. Summative assessments of speaking skills in the classroom bring their own set of challenges not only due to the time-consuming nature of these activi-

1 Gabriela Lüthi, Universität Freiburg, Institut für Mehrsprachigkeit, Murtengasse 24, 1700 Freiburg, Schweiz, E-Mail: gabriela.luethi@unifr.ch

2 Dr. Elisabeth Peyer, Universität Freiburg, Institut für Mehrsprachigkeit, Murtengasse 24, 1700 Freiburg, Schweiz, E-Mail: elisabeth.peyer@unifr.ch

3 Nadia Ravazzini, Universität Freiburg, Institut für Mehrsprachigkeit, Murtengasse 24, 1700 Freiburg, Schweiz, E-Mail: n.ravazzini@unibas.ch

ties, but also because they present teachers with a particularly complex task. In particular, the setting demands that teachers simultaneously take on multiple roles – those of test administrator, assessor and, depending on the test format, interlocuter – that in standardised international tests are generally distributed across several people. Thus, many teachers find speaking skills particularly difficult to assess (e. g. Ericksson & Gustafsson, 2005). For learners, foreign language speaking tests are often anxiety-inducing (Huang, 2018).

Against this background, the present study is interested in how students and their teachers of German as a foreign language in lower secondary school experience and view classroom-based speaking tests. Do they find them useful or rather a source of stress? Are the students aware of the learning objectives and do they know how to prepare for tests? And what views and wishes do the students have regarding the feedback they receive after a test? In the setting chosen for the study – the German as a foreign language classroom at the lower secondary level in French- and Italian-speaking Switzerland – these questions are of particular importance, as German is a compulsory school subject and a student's German marks can have a decisive impact on their future school trajectory.

2 Students' Views Regarding Language Assessment

In recent years, a large number of studies have been conducted on teachers' language assessment literacy, usually investigating (English as a foreign language) teachers' practices, views and training needs regarding foreign language assessment (e. g. Fulcher, 2012; Vogt et al., 2018; Berry et al., 2019). However, there has been little research on other stakeholders, such as learners (see Gan & Lan's scoping study on language assessment literacy, 2022). This is surprising given that learners are directly affected by assessment and are therefore important stakeholders. As has repeatedly been criticised, learners are often seen as subjects whose performance is evaluated in the assessment process without their views being taken into account (e. g. Vogt et al., 2018; Butler et al., 2021; Butler, 2022). However, if assessment is not seen as separate from teaching and learning, but as interconnected (e. g. Black & Wiliam, 1998a; Lee & Butler, 2020), it is essential to pay attention to learners. As Lee and Butler (2020) emphasise, learners need a minimal understanding of language assessment in order for them to benefit from assessment for their learning. Watanabe (2011), who taught a course in assessment literacy to first-year university students, gives further reasons why knowledge of language assessment is important for learners: He sees teaching assessment literacy to students as a way to help them overcome test anxiety. Furthermore, assessment literacy should also help to actively involve learners in the assessment process and thus increase their motivation.

The aim of our study was not to investigate students' language assessment literacy per se, but to understand how they perceive and experience classroom-based speaking tests. As Butler (2022) showed in an interview study with primary school EFL students,

even young learners can express their views about foreign language assessment. For instance, they can explain why they would like to see less form-focused assessment and how assessment practices could be improved. Similarly, an interview study by Czura (2017) with adolescent EFL learners showed that they were able to discuss the reliability and authenticity of an exam. Butler (2022) concludes from her study that considering learners' experiences and views can provide valuable information for improving assessment practices and make language assessment literacy more directly connected to learning and instruction. In addition, possible discrepancies between teachers' and students' understanding of the target abilities being assessed may be narrowed (Butler, 2022). Such differences in students' and teachers' perceptions of assessment were found in studies by Tarnanen and Huhta (2012), Vogt et al. (2020), Vlanti (2012) and Sato and Ikeda (2015). Feedback, in particular, seems to be an aspect of assessment where students' and teachers' perceptions often differ. For instance, in Vogt et al.'s (2020) large-scale survey of almost 1800 EFL learners and their teachers, teachers reported providing feedback in the form of brief or detailed comments and hints on how to improve learning more frequently than learners reported receiving such feedback. Tarnanen and Huhta's (2012) representative survey of around 1700 students and their teachers also revealed differences in the perception of feedback practices: the teachers' estimates of the frequency of different types of feedback usually were higher than the students', with the greatest discrepancy found in oral feedback to individual learners. In her survey on assessment practices in Greek EFL classrooms, Vlanti (2012) found that significantly more teachers than high school students claimed test activities to be similar to the activities done in the classroom. Sato and Ikeda (2015) discovered discrepancies between university students' perception of the ability being measured by items in high-stake tests and test developers' intentions. For example, items developed to test writing skills were interpreted as reading items. The studies by Vlanti (2012) and Sato and Ikeda (2015) show the importance of transparent communication of the purpose, aim and format of a test during exam preparation, with teachers playing a crucial role (see also Lee & Butler, 2020).

Our study aims to investigate secondary school students' views on different aspects of classroom-based speaking tests, with the intention of contributing to making these tests as useful and stress-free as possible for learners. The following aspects were examined by means of a written survey: students' perceptions of the exam preparation, the usefulness of speaking tests and the feedback they receive as well as the anxiety they feel during various test formats. In order to have a concrete point of reference, a group of students was also interviewed directly after having taken a speaking test. These students answered questions regarding exam preparation and the usefulness of speaking tests. Furthermore, the students' teachers were interviewed about exam preparation, feedback and usefulness of speaking tests. Wherever possible, the teachers' answers are compared to those of the students in order to reveal the extent of agreement or disagreement between the two groups' views.

3 The Context

This paper reports on part of a larger study that investigated classroom-based speaking assessment in German as a foreign language in lower secondary school in Switzerland (Peyer et al., 2025; Peyer et al., in press). In this study, authentic speaking tests, i.e. tests developed and administered by the teachers, were filmed. In addition to the filmed tests, students and teachers were interviewed and students also completed a written questionnaire. This paper focuses on data from the student questionnaire and student interviews and will be completed by the teachers' views expressed in their interviews.

In French-speaking and Italian-speaking Switzerland, German as a foreign language is a compulsory subject. In French-speaking Switzerland, primary school children start learning German in 3rd grade, and the minimum curricular learning outcome for oral production at the end of compulsory schooling in 9th grade is A2.2 (Conférence Interkantonal de l'Instruction Publique, 2012). In Italian-speaking Ticino students start learning German in 7th grade, the minimum learning outcome at the end of 9th grade is A1.2 (Dipartimento dell'educazione, della cultura e dello sport [DECS], 2015).⁴

Within French-speaking Switzerland, the cantons may have their own specific guidelines, which is the case for guidelines on assessment. In the Canton of Fribourg, two to three or three to four (graded) speaking tests are mandatory per year, depending on the number of weekly lessons (Service de l'enseignement obligatoire de langue française, 2020). By contrast, in the Canton of Neuchâtel speaking tests are merely recommended (Direction de l'instruction publique du canton de Berne et al., 2019). In Italian-speaking Ticino, speaking tests were also recommended at the time of data collection (DECS, 2015).⁵

As to students' motivation, the general public discourse is that German is an unpopular subject. This is reflected by the students in our sample: about half of them (somewhat) disagree with the statement "I like German".⁶ Notably, while students do not particularly like German, many do think that they are (somewhat) likely to need it later in their lives.

4 An updated version of the curriculum (September 2022) now only gives the higher level A2.1 as the learning objective for the speaking skill in German (DECS, 2022, p. 108).

5 The updated version of the curriculum, dated September 2022 (after our data collection), now requires teachers to test all skills and "not only grammatical or lexical elements" (DECS, 2022, p. 116). Even before this change to the curriculum, the cantonal pedagogic experts told teachers that testing all skills equally was important; this guideline was apparently also recorded in an unpublished internal document (P. Pfeifhofer, personal communication, 24.11.2023).

6 For comparison: in a representative survey of 2,000 secondary school students in central Switzerland, students were even more critical of compulsory French as a foreign language: only 33.3% indicated that they liked or rather liked learning French (while 69.6% indicated that they (rather) liked learning English) (Peyer et al., 2016).

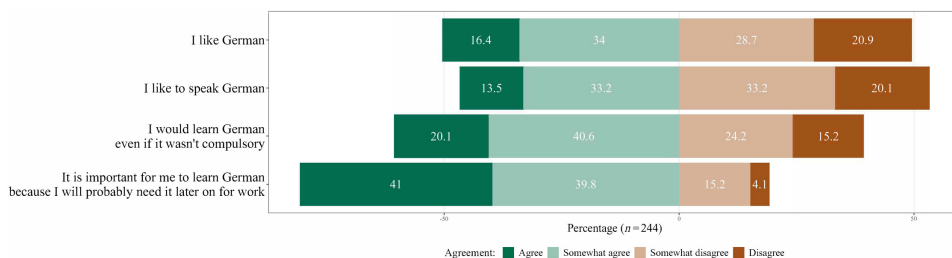


Figure 1: Results of the questionnaire on students' motivation (Source: own illustration)

4 The Study

4.1 Purpose and Research Questions

The purpose of our study was to understand how students view the speaking tests they regularly have to take in order to receive information for improving assessment practice. To ensure that speaking tests are meaningful for learners, it is important that students are able to prepare in a targeted manner and know the learning objectives. Furthermore, test results should be reported to learners in a way that they can identify their strengths and weaknesses. As Black and William (1998b) emphasise, feedback on tests should give “guidance on how to improve, and each pupil must be given help and an opportunity to work on the improvement” (p. 10). As oral exams are often associated with stress and anxiety (e.g. Hewitt & Stephenson, 2012; Butler et al., 2021) and test anxiety can have a negative impact on test achievement (e.g. Chapell et al., 2005), we also asked students about this point in order to obtain information about the conditions under which learners are least stressed.

The following research questions were asked:

1. How useful do students find the different formats of speaking tests for improving their German skills?
2. Are students aware of the learning objectives, and do they know how to prepare for the tests? How do the students prepare for a speaking test?
3. What feedback do students receive after a speaking test? What are their views and wishes regarding the feedback on speaking tests?
4. What are the student's anxiety levels regarding the different formats of speaking tests?

To explore these questions, a mixed method design was selected: on the one hand, a questionnaire was used to survey a large number of students about their views on speaking tests; on the other hand, retrospective interviews were used to gather students' views on a specific speaking test, i.e. the filmed test. In this way, students' general views on speaking tests could be supplemented with their views on a specific test, thus counteracting one disadvantage of written surveys, namely that respondents can-

not always accurately remember the situations asked about (e.g. Jones et al., 2013). It furthermore allowed students to provide more in-depth answers than in the more superficial questionnaire (Dörnyei & Taguchi, 2010).

4.2 Participants

The questionnaire was filled in by 254 lower secondary students, 187 from the French-speaking Cantons of Fribourg and Neuchâtel and 76 from Italian-speaking Ticino. Table 1 gives an overview of the number of students per track and per grade. The large majority of students spoke the language of schooling (as one of their languages) at home, i. e. either French (84 % of the students in the French-speaking cantons) or Italian (97 % in Ticino). In addition to the language of schooling, 33 % of all students reported speaking at least one more language at home. In some cases, this additional language was German (ten students, 4 %) or Swiss German (three students, 1.2 %). Most students (93 %) started learning German in either 3rd or 7th grade, according to the respective curriculum, and have had German lessons ever since.

The short interview was conducted with the 84 students whose speaking tests were filmed, 62 from the French-speaking region and 22 from Ticino. All but one of the interviewed students also completed the written questionnaire. Compared to the questionnaire, the interviewed students were slightly older and there were more students from lower tracks.

Table 1: Participants of the questionnaire and the interviews

Data	<i>n</i>	Track	Grade	Age
Questionnaire	254 (17 classes, 11 teachers)	Higher: 78 Middle: 86 Lower: 77 No track ⁷ : 13	7 th : 33 8 th : 114 9 th : 108	14.2 (12–17)
Interviews	84 (8 classes, 7 teachers)	Higher: 15 Middle: 14 Lower: 43 No track: 11	7 th : 12 8 th : 48 9 th : 24	15.8 (12–17)

4.3 Questionnaire

The questionnaire was created by the authors, with a few questions based on other questionnaires (e.g. Peyer et al., 2016; Vogt et al., 2018). During the development phase, the questionnaire was discussed with an expert. Once the questionnaire was finished, the German working document was translated simultaneously into French and Italian. The aim was that the French and Italian questionnaires were as similar as

⁷ In the canton of Neuchâtel, German is taught without differentiation until the end of 7th grade. It is in 8th and 9th grade that there is a higher and a lower track.

possible. For this reason, a departure from the German working document was at times necessary.

The initial questionnaire was a four-page pen-and-paper questionnaire that comprised a) five questions on background information, b) 37 statements to be answered with a 4-point Likert-type scale of agreement with one final open question and c) self-assessment with can-do descriptors from *Lingualevel* (Lenz & Studer, 2009) in the A1.2-B1.1 range. This initial questionnaire was piloted qualitatively in a 7th grade class in French-speaking Switzerland. For this purpose, 13 students filled in the questionnaire while sitting next to a researcher. After every set of questions, the students were asked if they understood everything or if anything was unclear or oddly worded. The students took about 8–12 minutes to complete the questionnaire. Overall, students reported almost no difficulty with the questionnaire. After the piloting, a few items were reformulated, and a few statements were exchanged for others or replaced by an open question.

The definitive questionnaire is an anonymous pen-and-paper questionnaire that was administered in the language of schooling, i.e. in French or Italian.⁸ It is four pages long, with one additional page for a self-assessment. It comprises three parts. The first part contains five questions on background information (age, gender, languages). The second and main part consists of six sets of statements, sometimes with an open question, on topics such as the general appreciation of German, preparation for exams, feelings of anxiety during exams, the usefulness of exams and feedback. In total, this part consists of 30 statements and four open questions. The statements are to be answered with a 4-point forced Likert-type scale, ranging from “agree”, to “somewhat agree”, to “somewhat disagree”, to “disagree”. Two sets of questions, pertaining to stress and usefulness of different test formats, also include the option “never done such an exam”. The third and final part is one page for self-assessment with can-do descriptors in the A1.2-B1.1 range for the French-speaking students or in the A1.1-A2.2 range for the Italian-speaking students – this self-assessment was completed by the filmed students.

4.4 Student Interviews

The students gave a short, approximately five minutes long semi-structured interview immediately following their speaking tests. The interviews were conducted in the language of schooling. The student interview was piloted with 13 students from the same class that piloted the questionnaire. Following the piloting of the interview, the questions were revised. The final student interview comprised questions on the students’ awareness of or feelings towards the camera, self-evaluation, preparation for the speaking test, usefulness of speaking tests and, if time permitted, a question on what they would change about the test.

8 All survey instruments are available here: <https://osf.io/yv8z9>.

4.5 Teacher Interviews

The semi-structured retrospective interviews with teachers took place after the speaking tests. They were conducted online via MS Teams either on the day of or the day after the speaking test. The interviews lasted about an hour and covered the following three phases: a) the exam preparation and the task, b) the rating and c) the exam follow-up (assessment for learning, feedback). In this paper, we include the teacher's answers about exam preparation and the exam follow-up.

4.6 Data Collection

The participants were chosen by convenience sampling. The classes' teachers were recruited by cantonal authorities or by the researchers' professional network, and their participation was voluntary. The students' participation was also voluntary: parental written consent was required for filming the speaking tests and giving the short interviews and/or for filling in the questionnaire; students who were 16 or older could also give consent themselves to fill in the questionnaire. On average, 65 % of students per class were recorded and 80 % completed the questionnaire.

Data collection mostly took place between March and June 2022, with two more classes filling in the questionnaire in the early spring of 2023. Speaking tests were filmed in eight classes. While tests were being filmed, members of the research team conducted the short interviews with the filmed students – students were interviewed in pairs after paired speaking tests (73 students) or alone after individual speaking tests (11 students) – or distributed the questionnaire. The students filled in the questionnaire individually, which took them 10–15 minutes. In addition to the eight classes in which speaking tests were filmed, the questionnaire was completed by students in nine further classes.

4.7 Data Analysis

The questionnaire data was manually entered and coded. The coding of the closed items was for the most part straight forward. The only time the research team made a decision on the coding was when students checked both “agree”/“somewhat agree” or “disagree”/“somewhat disagree”. In those rare instances, the “somewhat (dis)agree” option was coded. The questionnaire was fully completed by 218 students (86 % of questionnaires); in most cases, only one or two items are missing. No student was excluded from the analysis, however only students who answered all items of a set of questions were included in the analysis of that set. The quantitative data was analysed with descriptive statistics, the graphics were made with tidyverse (Wickham et al., 2019) in R (R Core Team, 2023). As the sample of participants was not representative and the overall study followed an explorative approach, no hypotheses were established, and consequently no inferential statistics calculated. The qualitative data, i. e. the answers to the open questions in the questionnaire and the interviews with students and teachers, were first transcribed and then examined following Mayring's (2015) approach to content analysis, specifically frequency analysis.

5 Results

The results of the study are presented below, organised thematically according to the four research questions, i. e. the students’ views on different aspects of speaking tests (usefulness, preparation, feedback and anxiety) are discussed and contrasted with the teachers’ views.

5.1 Usefulness

In the questionnaire, students indicated their agreement with statements on different test formats helping them to improve their German. As figure 2 shows, students agreed most that the format of paired speaking tests helps them improve. However, the students were overall positive about the different test formats being useful for their progress in German. This positivity is notable, as the learners in the European TALE project found oral presentations less conducive to language learning than our students (Vogt et al., 2020).

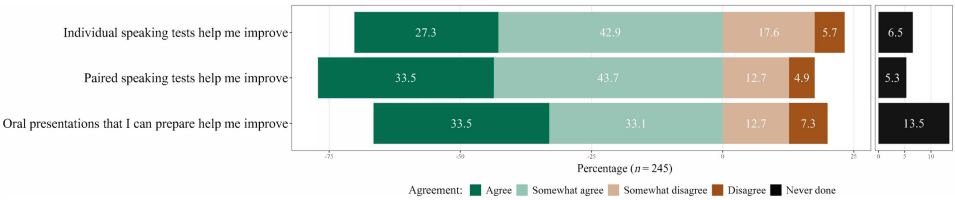


Figure 2: Results of the questionnaire on the usefulness of test formats (“... help me to improve my German”) (Source: own illustration)

The students whose speaking test was filmed and who were subsequently interviewed (n = 84), were asked if they thought speaking tests helped them improve. Here too, many students assented, with 57 % of the interviewed students finding them helpful and 23 % finding them at least partially helpful, while 18 % of the students did not find them helpful. Unlike the questionnaire, the interview allowed for students to explain their opinions. Their reasons for finding speaking tests (somewhat) helpful generally pertained to speaking in and of itself or to the usefulness of speaking compared to writing. The most common argument (17 %) was that they could learn or practice speaking with speaking tests. The second most common argument (12 %) was that they would need to speak German later in life (e. g. apprenticeship, work) or when travelling in a German speaking country – presumably, speaking tests help them prepare for this. The third most common argument (10 %) was that speaking tests allowed students to practice pronunciation. Other arguments that were named by multiple students (each 6 %) were first that speaking was the most fundamental part of the language, or at least more important than writing; second that speaking tests were not just about memorising, students had to work with what they knew; third that speaking tests helped students learn or review vocabulary. Students who did not find speaking tests useful had a variety of arguments against them, the most common of which was related to stress

(7%). Two further reasons were given by more than one student: the first was that they memorised things for the test they either forgot afterwards or did not understand (4%), or that the speaking tests were invented, inauthentic situations (2%).

Many of the students' arguments in favour of the usefulness of speaking tests seem to boil down to a) speaking tests are an opportunity to practice speaking or aspects of speaking, or b) speaking is important and useful, therefore speaking tests are useful. In the interviews with 12 students, after a student gave this kind of answer, one interviewer asked if speaking activities in the classroom would not be just as effective as speaking tests. Only three students maintained that speaking tests had an added benefit, either because the students stated that they spoke more during speaking tests than in classroom activities or because they considered being assessed to be a good thing. The other nine students concurred that speaking tests and classroom activities were similarly useful, with one student insisting on the importance of feedback: if she were to get feedback from the teacher in a classroom speaking activity (in her case on pronunciation), as she apparently did in the speaking test, then it would not make a difference.

The seven teachers also expressed their opinion on the usefulness of speaking tests. When asked whether they believed that speaking tests helped their students to progress in German, they expressed mixed opinions. One teacher considered speaking tests to be mainly "a source of stress and not a source of progress" for the students – she thought that other things, such as meeting a partner class from German-speaking Switzerland, would be more likely to motivate students to speak German. She also doubted that speaking tests help her students because of the inauthentic situation and because students do not make use of the feedback they receive, in her case a filled-out grid with an occasional comment. A few other teachers were also doubtful that the students used the feedback, however, they still found speaking tests useful for other reasons. Another teacher, who participated with two classes, and was thus interviewed twice, was doubtful in the first interview if the class in question (higher track, 8th grade) profited from speaking tests because the students at that age would not yet see the point in learning German. In the second interview (lower track, 9th grade), the teacher expressed her hope that students would benefit from speaking tests, adding that she thought it was important for the students to have had this experience. Four teachers viewed speaking tests as a means of pressure or external source of motivation, saying that speaking tests helped students to progress because students studied more for tests than for regular class. However, two of these teachers added the caveat that this was only true for higher track German classes and not the lower track classes with which they participated. The other two teachers, who also participated with lower track classes, maintained that speaking tests helped these students to progress because students are more "motivated" and "put in more effort" or "study more seriously". The final teacher argued differently: he thinks that speaking tests are more useful than any other type of test. For one thing, they allow the students to become aware of their level, specifically if they could or could not interact in an everyday situation. Consequently, he expected students either to feel assured or to realise that they have to work harder to

reach the required level. For another thing, he could give students immediate feedback or even help them during the test.

5.2 Preparation

One of the teachers' arguments in favour of the usefulness of speaking tests – and that of a few students as well – was that students studied more before a test. In other words, the preparation for the test was conducive to students learning. In the interviews, the teachers were asked about the preparation for the (filmed) test. The teachers prepared their classes more or less intensely for the speaking tests. On the one side, one teacher did a mock test with her class. On the other side, one teacher used parts of two lessons to not only familiarise the students with the test task but also review the topic of the test (describing a daily schedule) which had been covered some months earlier.

In the questionnaire, the students indicated that their teachers informed them of the topic and learning objectives of the speaking exams in class. In fact, this was the item with the highest agreement in the questionnaire. Though students definitely agreed that they had been told the learning objectives by their teachers, students were less likely to know them and even less sure about how to prepare for speaking tests.

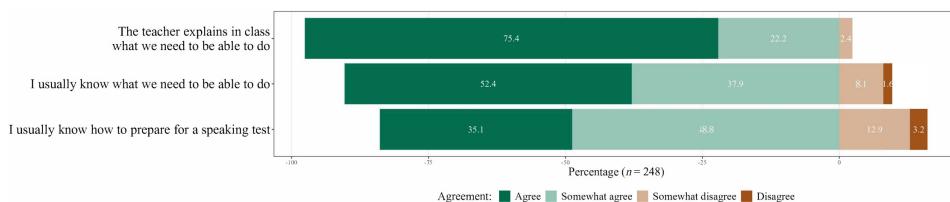


Figure 3: Results of the questionnaire on how students know the learning goals for speaking tests (Source: own illustration)

The next question in the survey was how students prepare for speaking tests in German. Since this was an open question, multiple answers were possible. Of the 244 students who answered the question, 32 % students said that they studied or reviewed vocabulary. 19 % of the students indicated that they studied with a family member, usually with their parents. 17 % each mentioned either revision without stating what exactly they reviewed or reviewing and preparing sentences or questions that they expect to say during the speaking test. 12 % of the students prepare by writing a text, basically a script of their test talk, and some of them also explicitly stated that they also memorise the written text. Further answers given by multiple students were that they practiced speaking, e. g. by speaking aloud (8 %), that they memorise things (8 %) – it is often unclear what exactly students memorise, but in many cases it seems to be a document they received from the teacher, perhaps containing the most important phrases for the test – that they read the learning objectives (7 %), that they (re)do exercises, e. g. in the coursebook (6 %), they study with friends/classmates (5 %) or another unidentified person (4 %), they focus on practicing their pronunciation (4 %), or they refer to the

preparation done in the classroom (4%). 6 % of the students wrote that they do nothing to prepare for speaking tests.

In the short interview, the students ($n = 84$) were asked how they had prepared for the (filmed) test they had just passed. There is less variety in the answers in the interviews than in the questionnaire. In the interviews, the most common answer was that they had reviewed sentences or questions that they would or could use during the exam (32 %). The second most common answer was that they reviewed the vocabulary (30 %). Students also stated that they had studied with family members, mostly their parents (21 %) or with a friend/classmate (19 %).

5.3 Feedback

For learners to benefit from feedback on tests, they need to receive more than just a grade. Indeed, according to Black and Wiliam (1998b), feedback should provide students with information on how to improve. In the interviews, the teachers talked about the feedback they were going to give their students after the speaking test. All teachers gave them their grade. Five of the seven teachers also gave the students the filled-in grid, which may contain a written comment or two (e. g. a mistake). There was however doubt among teachers, based on their experience, that students (know how to) use the feedback for further learning, with one teacher mentioning that this was less of an issue in the higher track classes. Three teachers offered the students the opportunity to ask questions about the test or the grade. One teacher always briefly talks to every student individually to explain the points in the grid and the grade. Here too, some teachers said that students are mainly interested in getting their grade and if they ask questions after the test, it is to understand why they did not get more points or to argue for a better grade. Other feedback that teachers offered their students was the opportunity to listen to the audio recording of the test, however students never took the teacher up on this. Another teacher said in the interview that he would try a new way of giving feedback: he planned on listening to the recording with the students and then giving them feedback. Furthermore, one teacher always gave the entire class feedback by commenting on a few mistakes she had heard during the tests. Two teachers also took note of a specific grammar point that they considered to be important and where they frequently observed students having made errors; they would then work on this grammar point in the weeks following the test or during the end-of-year revision. Two teachers also mentioned that they use the feedback to encourage anxious students or the entire class.

In the questionnaire, students were asked what type of feedback they would like to receive after speaking tests; they were also asked to motivate their answers. As figure 4 shows, students indicated that they would like to receive all sorts of feedback, with tips on how to improve being the most popular. The students' reasons for wanting any or all types of feedback were related to using the feedback to improve. The least favoured amongst the types of feedback was the teacher's feedback on the entire class's performance, some students arguing that they did not care or need to know about the other students' performances.

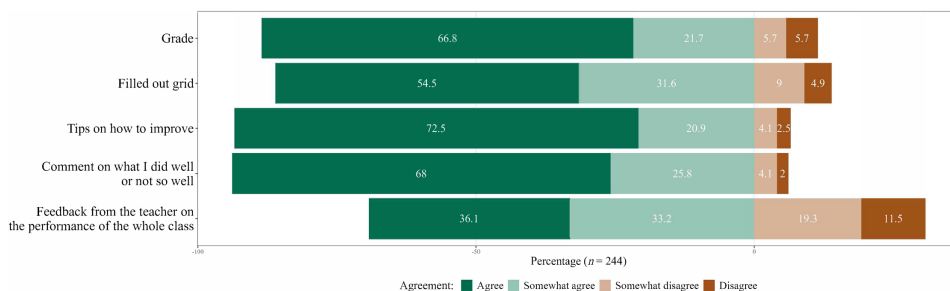


Figure 4: Type of feedback students would like to receive after speaking tests (Source: own illustration)

In a further, open question, the students were asked what they did with the feedback they received from their teachers. Multiple answers were possible. When analysing these responses, it is important to consider that not all students receive the same kind or the same amount of feedback. Nevertheless, their answers (n = 244) fall into three categories⁹: a) using the feedback for future learning (57 %), b) looking at the feedback (30 %), c) not using the feedback (19 %). Within the first category of students using the feedback for future learning, there were different degrees of specificity in the answers. Many gave vague answers such as “I improve”; some students said that they would learn from mistakes or other things they did not do so well in the test; fewer students mention putting in work to improve, e.g. “I look at where I have to improve the most and I practice at home and try to improve”. Finally, some students indicated that they used the feedback for the (preparation of) the next speaking test. Within the second category of students looking at the feedback, there were also different degrees of specificity. Here, most say that they look at mistakes and weaknesses – only three students also looked at the positive points, i.e. at what went well or their strengths – with some other students simply indicating that they look at the feedback without any further details. The last category of students, who do not use the feedback, either answered directly that they do nothing with the feedback or gave an answer that made clear that they do not use the feedback for further learning, e.g. putting it in a binder, showing their parents or saying that they are happy if it is positive feedback and disappointed if it is negative feedback. Overall, the responses suggest that many students understand that feedback can promote further learning. However, the often very vague statements also suggest that the students rarely (are able to) incorporate the feedback into their learning, thus also raising the question as to how actionable teacher feedback is for the students.

5.4 Anxiety

In the questionnaire, the students were specifically asked about their feelings of anxiety in different test formats. As figure 5 (line 3–7) shows, students overall report feeling anxious or stressed¹⁰ during speaking tests. Of the two test formats that students were

⁹ Some answers were assigned both to the first category (use for future learning) and to the second (look at feedback).

¹⁰ The French and Italian questionnaire used the word “stressé-e” and “stressato/a” to describe a feeling of test anxiety.

asked about, individual or paired tests, students feel less anxiety during paired speaking tests – as mentioned above, this was also the test format students found the most useful. Other factors that may influence feelings of anxiety were also included in the questionnaire. Of these, students were most anxious when taking the test in the classroom with the other students present. Most students in our sample have some experience in taking a test in the classroom, as this is how the teachers in our study usually administer speaking tests, for different reasons: classroom management, i. e. to make sure students stay in the room and behave, as well as a kind of preparation for the end of school exam that takes place in a room with other students also taking the test, or practicing presentation skills and speaking in front of people. Another factor that was likely to stress students was not having any preparation time; fortunately for the students, in the (filmed) tests, they were accorded some time before the test. It is interesting to note that a comparison between students of the lower and the other tracks shows that the lower track students tended to report less anxiety while speaking in class or during tests. Furthermore, a comparison between male and female students revealed a tendency for girls to be more anxious than boys in all of the assessment situations mentioned in figure 5. This is in line with previous studies (Ericksson & Gustafsson, 2005; Gursoy & Arman, 2016). In Ericksson and Gustafsson's (2005) study, girls indicated more frequently to feel nervous, most of all in formal testing situations, but also in classroom assessment.

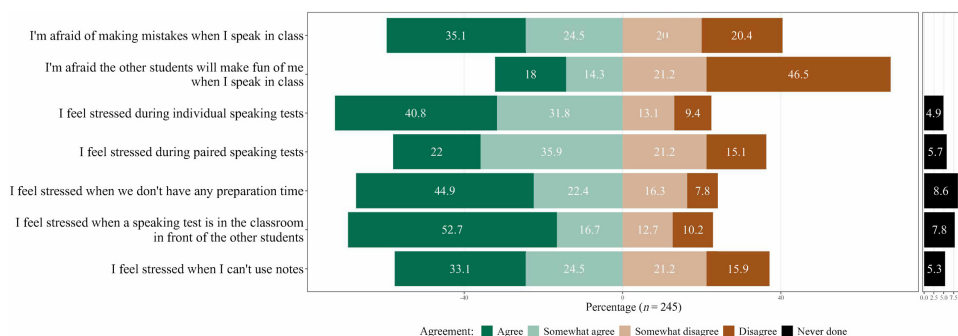


Figure 5: Reported test anxiety (Source: own illustration)

Anxiety is clearly an issue on students' minds when it comes to speaking tests. In the last question of the questionnaire, students were asked what else they would like to say about speaking tests in German. 43 (17%) of all 254 students spontaneously wrote how stressful they find speaking tests. While students did not provide further details, five students indicated that they find it particularly stressful when speaking tests take place in the classroom with the other students present.

6 Discussion

This study investigated students' views on German as a foreign language tests at lower secondary level. 254 learners completed a questionnaire on exam preparation, usefulness, feedback and anxiety, and 84 learners were interviewed immediately after a speaking test. In order to identify possible differences in the perceptions of teachers and learners, their teachers ($n = 7$) were also interviewed. It should be kept in mind that this is a non-representative convenience sample, as the number of participants (especially participating classes) is limited and that participants took part in the study on a voluntary basis. Thus, the participating students are most likely taught by teachers who had a measure of confidence in the assessment of speaking. For the above reasons, the results cannot be generalised, and further research is necessary to solidify the results. Nonetheless, we believe that the results reveal valid tendencies. The results thus provide information about current views and practices and serve as a basis for further discussion and research.

In the following, the results of the present study will be discussed with reference to the research questions, and possible discrepancies between students' and teachers' views are pointed out. The first research question pertained to how useful students find various speaking test formats. The analysis of the students' questionnaire showed that two thirds of the learners have positive attitudes towards speaking tests, with paired speaking tests being considered the most useful test format – 77 % of the students say they (tend to) benefit from them. Students who were interviewed directly after a speaking test also found speaking tests (rather) useful for progressing in German in 80 % of cases. The interviewed students explained their positive opinions mainly by saying that speaking tests helped them to improve their speaking skills and that they needed to be able to speak German (later) in life. Hence, they mostly gave reasons for the importance of practicing speaking, rather than for speaking tests themselves. The learners in our study thus show similar views to those of Erickson and Gustafsson (2005), who appreciated language assessment that was applicable in daily life and often emphasised a wish for oral proficiency. The teachers expressed mixed opinions regarding the usefulness of speaking tests. Although a majority of the teachers views speaking tests as a means of getting students to study more, some specify that this is only true for higher track classes. Some teachers were also doubtful if students used the feedback they receive. The view of tests as a necessary external source of motivation was also expressed by some students as well as by learners of other studies (see Vavla & Gokaj, 2013; Agcam & Babanoglu, 2016). This perception of tests as an external source of motivation may point to shortcomings with German as a foreign language teaching in Switzerland. Although a compulsory subject from 3rd grade onwards, German is in a difficult position alongside the more popular English: pupils make only slow progress and lose their intrinsic motivation to learn it over the years.

As mentioned above, the majority of teachers in our study thought that the students study more before a test. This ties into the second research question of whether students are aware of the learning objectives and know how to prepare for a speaking

exam. Although the overwhelming majority of students indicated that they were generally informed of the learning goals by their teachers (75 % agreed, 22 % somewhat agreed), only half of the learners usually know the goals (52 %) and even fewer (35 %) usually know how to prepare for speaking tests. This difference between the ‘what’ (the learning objectives) and the ‘how’ (knowing how to prepare to fulfil them) is striking and a point that teachers should address in order to make speaking tests as useful and stress-free as possible for learners. After all, if students do not know how best to prepare for a test, it is likely that they study less and are perhaps also more nervous during the exam.

The third research question was concerned with the feedback students receive and would like to receive after speaking tests. For speaking tests to be an opportunity for learning, learners should be given feedback that contains concrete tips for improvement. When asked what kind of feedback they would like to receive, the students in our study indicated all sorts of feedback, with comments on how to improve being the most popular. The students also mostly explained their wish for any feedback with wanting to use it for improvement. However, the feedback that students – according to their teachers – receive after a test, rarely seems to contain concrete tips on how a student can improve. This discrepancy between students’ wishes and teachers’ practices is reminiscent of Vogt et al.’s (2020) and Tarnanen and Huhta’s (2012) studies, in which discrepancies between teachers’ and students’ perception of the frequency of individual feedback on how to improve learning were found. However, the results of our study not only show a discrepancy between students’ wishes and teachers’ practices but also regarding students’ self-perception and the teachers’ perception of them: some of the teachers in our study doubted that students are interested in using feedback to improve. They mention that students are mainly interested in getting their grade and usually only ask questions after the test to understand why they did not get more points. The teachers furthermore expressed doubts that students know how to use the feedback for further learning (for similar views of teachers, see also Tarnanen & Huhta, 2012). This could indicate that the students are either not motivated enough and/or do not have the ability and self-reflection to set themselves goals and consider how they can best be achieved. However, the analysis of the teacher interviews indicates that the feedback they give after the exams is often not individual, rather unspecific (i. e. not goal-oriented enough) and contains few concrete examples on how to improve.

Since oral exams can cause anxiety, which in turn can affect students’ performance (e. g. Hewitt & Stephenson, 2012), our fourth research question asked how stressed students feel during different formats of speaking tests. While a majority of students overall report feeling (somewhat) anxious during speaking tests, clear differences between the individual and paired formats were found: while 41 % of students are clearly stressed during individual tests, this proportion is only about half as high (22 %) for paired tests. This finding is in line with Fulcher’s study (1996) in which adolescent EFL learners were less anxious prior to a group discussion than prior to one-to-one interviews. However, the biggest stress factor, according to our study, is taking the test in front of the whole class: 53 % of students are clearly stressed in this setting. Teachers often stated during

the interviews that they conducted speaking tests in the classroom with the other students present in order to avoid disciplinary problems. However, given the large number of stressed students, it would certainly be desirable for teachers to use other exam settings more often, such as testing in a different room or having students make recordings of their discussions which teachers could evaluate later. Against the background of studies that show negative correlations between anxiety and oral test performance (e. g. Hewitt & Stephenson, 2012), it seems vital to help students to be relieved from anxiety and nervousness during oral tests. Furthermore, as Butler et al. (2021) emphasise, it would generally be important to pay more attention to the role of affect and its influence on language assessment.

7 Conclusion

The aim of this study was to shed light on an under-researched aspect of language assessment: the views of learners as arguably the most important stakeholders. By combining quantitative and qualitative research methods, we hope to have shown that it is worthwhile considering the opinions and wishes of secondary school students. Although the results of our study cannot be generalised to other contexts due to non-representative convenience sampling, they can nevertheless provide indications of how teachers can make speaking tests as anxiety-free and learner-oriented as possible. The discrepancies found between teachers' and learners' views furthermore point to a need for more communication about assessment between teachers and learners, which should hopefully lead to students knowing better how to prepare for a speaking exam. The results also point to a need for teacher training and professional development especially in providing learner-oriented feedback before or after speaking tests. Further research into students' perspectives on speaking tests is necessary to bolster and add to our findings. There are still many open questions about how students, especially teenagers in compulsory education, experience speaking tests. It would, for example, be interesting to research if students' views remain somewhat stable or if with more experience, i. e. after more speaking tests, their views may change. With a large-scale representative sample, the question of teacher or class effects on students' views could also be addressed. Combined with teacher profiles, this may provide best practices for teachers and their classroom with regard to speaking tests, which could contribute to making speaking tests as anxiety-free and useful as possible for students.

References

- Agcam, R., & Babanoglu, M. P. (2016). Students' Perceptions of Language Testing and Assessment in Higher Education. *Üniversitepark Bülten*, 5(1–2), 66–77. <https://doi.org/10.22521/unibulletin.2016.512.6>

- Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal*, 73(2), 113–123. <https://doi.org/10.1093/elt/ccy055>
- Black, P., & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Black, P., & Wiliam, D. (1998b). Inside the Black Box: Raising Standards through Classroom Assessment. *Phi Delta Kappan*, 92(1), 81–90. <https://doi.org/10.1177/003172171009200119>
- Butler, Y. G. (2022). Exploring the Idea of “Assessment with Learners” in Discussing Language Assessment Literacy. *Language Teaching Research Quarterly*, 29, 92–101. <https://doi.org/10.32038/ltrq.2022.29.06>
- Butler, Y. G., Peng, X., & Lee, J. (2021). Young learners’ voices: Towards a learner-centered approach to understanding language assessment literacy. *Language Testing*, 38(3), 429–455. <https://doi.org/10.1177/0265532221992274>
- Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test Anxiety and Academic Performance in Undergraduate and Graduate Students. *Journal of Educational Psychology*, 97(2), 268–274. <https://doi.org/10.1037/0022-0663.97.2.268>
- Conférence Intercantonale de l’Instruction Publique. (2012). *Plan d’études romand. Commentaires généraux du domaine langues*. <https://www.plandetudes.ch/web/guest/1/cg>
- Czura, A. (2017). Adolescent learner perceptions of foreign language assessment: Critical incident analysis. *Glottodidactica*, 44(2), 25–39. <https://doi.org/10.14746/gl.2017.44.2.02>
- Dipartimento dell’educazione, della cultura e dello sport, Repubblica e Cantone Ticino. (2015). *Piano di Studio della scuola dell’obbligo ticinese*. https://m4.ti.ch/fileadmin/DECS/DS/documenti/pubblicazioni/ricerca_educativa/2024_arch_Piano_di_studio_della_scuola_dell_obbligo_ticinese_-_VERSIONE_2015.pdf
- Dipartimento dell’educazione, della cultura e dello sport, Repubblica e Cantone Ticino. (2022). *Piano di Studio della scuola dell’obbligo ticinese*. <https://pianodistudio.edu.ti.ch/wp-content/uploads/2024/01/Piano-di-studio-perfezionato.pdf>
- Direction de l’instruction publique du canton de Berne ; Département de la formation de la culture et des sports de la République et du Canton du Jura & Département de l’éducation et de la famille de la République et du Canton de Neuchâtel (2019). *Allemand aux cycles 2 et 3: Document tri-cantonal (BEJUNE) du 5 septembre 2019 (évolutif)*.
- Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: construction, administration, and processing* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203864739>
- Erickson, G., & Gustafsson, J.-E. (2005). *Some European students’ and teachers’ views on language testing and assessment: A report on a questionnaire survey*. European Association of Language Testing and Assessment.
- Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing*, 13(1), 23–51. <https://doi.org/10.1177/026553229601300103>
- Fulcher, G. (2012). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly*, 9(2), 113–132. <https://doi.org/10.1080/15434303.2011.642041>

- Gan, L., & Lam, R. (2022). A Review on Language Assessment Literacy: Trends, Foci and Contributions. *Language Assessment Quarterly*, 19(5), 503–525. <https://doi.org/10.1080/15434303.2022.2128802>
- Gursoy, E., & Arman, T. (2016.) Analyzing Foreign Language Test Anxiety among High School Students in an EFL Context. *Journal of Education and Learning*, 5(4), 190–200. <http://dx.doi.org/10.5539/jel.v5n4p190>
- Hewitt, E., & Stephenson, J. (2012). Foreign Language Anxiety and Oral Exam Performance: A Replication of Phillips's MLJ Study. *The Modern Language Journal*, 96(2), 170–189. <https://doi.org/10.1111/j.1540-4781.2011.01174.x>
- Huang, H.-T. D. (2018). Modeling the relationships between anxieties and performance in second/foreign language speaking assessment. *Learning and Individual Differences*, 63, 44–56. <https://doi.org/10.1016/j.lindif.2018.03.002>
- Jones, T., Baxter, M., & Khanduja, V. (2013). A quick guide to survey research. *The Annals of The Royal College of Surgeons of England*, 95(1), 5–7. <https://doi.org/10.1308/003588413X13511609956372>
- Lee, J., & Butler, Y. G. (2020). Reconceptualizing Language Assessment Literacy: Where Are Language Learners? *TESOL Quarterly*, 54(4), 1098–1111. <https://doi.org/10.1002/tesq.576>
- Lenz, P., & Studer, T. (2009). *Handlungsbezogene Deskriptoren aus Lingualevel/action-oriented descriptors of Lingualevel*. Schulverlag. <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680492ff2>
- Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12th ed.). Beltz.
- Peyer, E., Andexlinger, M., & Kofler, K. (2016). *Projekt Fremdsprachenevaluation BKZ: Schlussbericht zu den Befragungen der Schülerinnen und Schüler*. Institut für Mehrsprachigkeit. <https://folia.unifr.ch/unifr/documents/304818>
- Peyer, E., Lüthi, G., & Ravazzini, N. (2025). *Beurteilung der Sprechkompetenz im DaF-Unterricht: Praktiken und Ansichten von Lehrpersonen und Schüler/innen der Sekundarstufe I / Assessment of speaking skills in the German as a foreign language classroom: Practices and views of lower secondary school teachers and students*. Institut für Mehrsprachigkeit. <https://folia.unifr.ch/unifr/documents/331231>
- Peyer, E., Lüthi, G., & Ravazzini, N. (in press). Teacher interventions following test takers' interactional troubles during classroom-based paired speaking tests at lower level. An exploration of (non) interventions and their effects on progressivity and test takers' independent interactional trouble solving. *Zeitschrift für Interaktionsforschung in DaFZ*.
- R Core Team (2023). R: A Language and Environment for Statistical Computing (version 4.3.0) [Computer software]. <https://www.R-project.org/>
- Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact of face validity on student learning. *Language Testing in Asia*, 5(1), Article 10. <https://doi.org/10.1186/s40468-015-0019-z>
- Service de l'enseignement obligatoire de langue française, Etat de Fribourg. (2020). *L'évaluation au CO: Compléments aux directives de juin 2006, version 2020*.

- Tarnanen, M., & Huhta, A. (2011). Foreign language assessment and feedback practices in Finland. In D. Tsagari, & I. Csépes (Eds.), *Classroom-based language assessment*. (pp. 129–146). Peter Lang.
- Vavla, L., & Gokaj, R. (2013). Learner's Perceptions of Assessment and Testing in EFL Classrooms in Albania. *Mediterranean Journal of Social Sciences*, 4(11), 509–515. <https://doi.org/10.5901/mjss.2013.v4n11p509>
- Vlanti, S. (2012). Assessment practices in the English language classroom of Greek Junior High School. *Research Papers in Language Teaching and Learning*, 3(1), 92–122. <https://rpltl.eap.gr/images/2012/03-01-092-Vlanti.pdf>
- Vogt, K., Brüstle, M., & Sperling, I. (2018). *TALE-Teachers' Assessment Literacy Enhancement: Needs analysis report*. <https://taleproject.eu>
- Vogt, K., Tsagari, D., Csépes, I., Green, A., & Sifakis, N. (2020). Linking Learners' Perspectives on Language Assessment Practices to Teachers' Assessment Literacy Enhancement (TALE): Insights from Four European Countries. *Language Assessment Quarterly*, 17(4), 410–433. <https://doi.org/10.1080/15434303.2020.1776714>
- Watanabe, Y. (2011). Teaching a course in assessment literacy to test takers: Its rationale, procedure, content and effectiveness. *Cambridge Research Notes*, 46, 29–34. <https://www.cambridgeenglish.org/Images/23164-research-notes-46.pdf>
- Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E. Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D., P., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 43(4), Article 1686. <https://doi.org/10.21105/joss.01686>

