pedocs
Open Access Erziehungswissenschaften

DIPF

Harsch, Claudia; Hartig, Johannes; Delgado-Osorio, Ximena; Koval, Valeriia

Erfassung und Modellierung integrierter akademisch-sprachlicher Kompetenzen (MASK). Abschlussbericht

Bremen ; Frankfurt 2026, 15 S.

Quellenangabe/ Reference:
Harsch, Claudia; Hartig, Johannes; Delgado-Osorio, Ximena; Koval, Valeriia: Erfassung und Modellierung integrierter akademisch-sprachlicher Kompetenzen (MASK). Abschlussbericht. Bremen ; Frankfurt 2026, 15 S. - URN: urn:nbn:de:0111-pedocs-365955 - DOI: 10.25656/01:36595

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-365955
https://doi.org/10.25656/01:36595

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: http://creativecommons.org/licenses/by-nc-nd/4.0/deed.de - Sie dürfen das Werk bzw. den Inhalt unter folgenden Bedingungen vervielfältigen, verbreiten und öffentlich zugänglich machen: Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen. Dieses Werk bzw. dieser Inhalt darf nicht für kommerzielle Zwecke verwendet werden und es darf nicht bearbeitet, abgewandelt oder in anderer Weise verändert werden.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License: http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en - You may copy, distribute and transmit, adapt or exhibit the work in the public as long as you attribute the work in the manner specified by the author or licensor. You are not allowed to make commercial use of the work or its contents. You are not allowed to alter, transform, or change this work in any other way.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

Abschlussbericht

Erfassung und Modellierung integrierter akademisch-sprachlicher Kompetenzen (MASK)

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn · Postanschrift: 53170 Bonn
Telefon: + 49 228 885-1 · Telefax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

ABSCHLUSSBERICHT

1 Allgemeine Angaben

DFG-Geschäftszeichen: HA 5943/2-1 respektive HA 5050/11-1 (MASK I), HA 5943/2-2 respektive HA 5050/11-2 (MASK II)

Projektnummer: 429641868

Titel des Projekts: Erfassung und Modellierung integrierter akademisch-sprachlicher Kompetenzen (MASK)

Name(n) der Antragsteller*innen:

Claudia Harsch, Universität Bremen

Johannes Hartig, DIPF I Leibniz-Institut für Bildungsforschung und Bildungsinformation

Dienstanschrift/en:

Universität Bremen, Fachbereich 10: Sprach- und Literaturwissenschaften
Universitäts-Boulevard 13, Gebäude GW 2, 28359 Bremen

DIPF I Leibniz-Institut für Bildungsforschung und Bildungsinformation
Rostocker Straße 6, 60323 Frankfurt

Name(n) der Mitverantwortlichen:

Valeriia Koval, Universität Bremen

Ximena Delgado Osorio, DIPF

Name(n) der Kooperationspartner*innen:

Prof. Torsten Zesch, Fernuniversität in Hagen

Prof. Pablo Pirnay-Dummer, Martin-Luther-Universität Halle-Wittenberg

Dr. Ulf Kröhne, DIPF I Leibniz-Institut für Bildungsforschung und Bildungsinformation

Dr. Voula Kanistra, Trinity College London

Sebastian Gombert, DIPF I Leibniz-Institut für Bildungsforschung und Bildungsinformation

Berichtszeitraum (gesamte Förderdauer):

01.2020 – 12.2022 (MASK I)

02.2024 – 12.2025 (MASK II)

2 Zusammenfassung / Summary

Das Projekt untersuchte die Kompetenz, integrierte akademische Schreibaufgaben in Englisch als Lingua Franca zu bewältigen – eine zentrale Fähigkeit im internationalen Hochschulstudium. Ziel war es, ein wissenschaftlich fundiertes Modell für diese Kompetenz zu entwickeln und zu validieren. In zwei Phasen (MASK I und II) wurden vier Themenbereiche bearbeitet: Kompetenzmodell und Diagnostik, Bewertungsprozesse, Bearbeitungsstrategien der Studierenden sowie computergestützte Textbewertung.

Im Projekt wurden authentische Aufgaben (Zusammenfassungen und Stellungnahmen) und ein detailliertes Bewertungsschema mit Kriterien wie Textnutzung, Diskurssynthese und Sprachqualität entwickelt. Mit 451 Teilnehmenden (Schüler*innen und Studierende) wurde die Leistung in diesen Aufgaben erfasst und mit isolierten Lese- und Schreibtests verglichen. Die Ergebnisse zeigen: Die Leistung in integrierten Aufgaben lässt sich nur teilweise durch isolierte Sprachfähigkeiten vorhersagen. Stattdessen ist die Fähigkeit, Quelltexte richtig zu nutzen, besonders wichtig – besonders bei Zusammenfassungen.

Interessant ist, dass auch unerfahrene Bewerter*innen mit angemessener Schulung hochwertige Bewertungen liefern können, während Expert*innen gelegentlich stärker von „Halo-Effekten“ beeinflusst sind. Die Bearbeitungsprozesse der Studierenden wurden mithilfe von Lautes-Denken-Protokollen und Logdaten analysiert: Studierende nutzen unterschiedliche Strategien je nach Aufgabentyp, und bestimmte Verhaltensmuster (z. B. gezielte Markierungen) korrelieren mit besserer Leistung.

Ein wichtiges Ergebnis ist die Entwicklung eines Tools zur Erkennung wortwörtlicher Textübernahmen, das bereits im Sprachunterricht in Bremen eingesetzt wird. Zudem zeigten KI-Modelle (BERT-basiert), dass automatisierte Bewertungen mit hoher Genauigkeit möglich sind – besonders bei Grammatik, Wortschatz und Zitierweise. Die Ergebnisse sind von großer Relevanz für Hochschulen: Sie können zu einer objektiverenn, effizientere und zugleich menschlich fundierten Bewertung von Schreibleistungen in internationalen Studiengängen beitragen.

This project investigated the competence required to successfully complete integrated academic writing tasks in English as a Lingua Franca – a crucial skill in international higher education. The goal was to empirically develop and validate a competence model. The research was conducted in two phases (MASK I and II), covering four key areas: competence modeling and diagnostics, rater processes, student writing processes, and computer-assisted automated scoring.

Authentic tasks (summarizing and responding to texts) were developed, along with a detailed scoring rubric assessing text use, discourse synthesis, and language quality. Data from 451 participants (high school students and first-year university students) showed that performance in integrated tasks is only weakly predictable from isolated reading and writing tests. Instead, the ability to accurately use source texts – especially in summaries – emerged as a key predictor.

Surprisingly, novice raters trained with clear guidelines achieved high-quality assessments comparable to experts, while experienced raters at points showed stronger halo effects. Student writing processes were analyzed using think-aloud protocols and log data (e.g., annotations, keystrokes). Results revealed distinct strategies depending on task type, with certain behaviors (like targeted highlighting) linked to higher-quality outputs.

A key innovation was a software tool to detect verbatim text copying, now used in the language center in Bremen. Furthermore, AI models (BERT-based) demonstrated strong potential for automated scoring, achieving high accuracy in predicting human ratings – especially

for grammar, vocabulary, and citation use. The lowest accuracy was found for complex synthesis tasks.

These findings are highly relevant for universities: they enable more reliable, efficient, and scalable assessment of academic writing in multilingual settings. The project's tools and insights can contribute towards fairer evaluation, better feedback, and a human-rater based Assessment of writing performances in international English-medium study programmes.

Ein Entwurf für die Zusammenfassung wurde mit Hilfe eines LLM (Quen 3 30B A3B Instruct 2507) auf academic-cloud.de erstellt.

3 Wissenschaftlicher Arbeits- und Ergebnisbericht

Das übergeordnete Ziel des Projekts war die empirische Fundierung des Kompetenzkonstrukts, das der erfolgreichen Bearbeitung integrierter akademisch-sprachlicher Aufgaben in einem Lingua-Franca-Kontext zugrunde liegt.

Das Projekt bestand aus zwei Phasen (MASK I und MASK II), deren Ziele vier Themenbereichen zugeordnet sind: (a) Kompetenzmodell & Diagnostik, (b) Bewertungsprozesse und Expertise-Effekte der Bewerter*innen, (c) Bearbeitungsprozesse der Studierenden und (d) Computergestützte & automatisierte Textbewertung.

a) Kompetenzmodell & Diagnostik: Theoretische Fundierung und empirische Validierung des Konstrukts integrierter Schreibkompetenzen.

In Bezug auf Kompetenzmodell und Diagnostik wurden im **MASK I** die folgenden Meilensteine erreicht: die *Entwicklung und Erprobung von Testaufgaben und Bewertungskriterien*, die *Prüfung der Konstrukt-Dimensionalität*, und die *Untersuchung vom Zusammenhang mit unabhängigen Lese- und Schreibkompetenzen und Studienerfolg*.

Als erstes wurden die Testaufgaben und das Bewertungsschema entwickelt (AP1). Dafür entwickelten zwei Hochschullehrerinnen und zwei Sprachlehrende am Sprachenzentrum vier integrierte Aufgaben, die sich an möglichst authentische Schreibsituationen anlehnen. Zwei Aufgaben erforderten eine Zusammenfassung eines authentischen Texts aus einem Einführungswerk, das sich an Studienanfänger*innen wendet, die anderen beiden Aufgaben verlangten eine Stellungnahme zu einem Einführungstext, der zwei gegensätzliche Darstellungen zu einem Thema enthält. Pro Aufgabentyp kamen jeweils ein Text zu einer naturwissenschaftlichen und zu einer gesellschaftswissenschaftlichen Thematik zum Einsatz. Die Ausgangstexte hatten eine Länge von ca. 1000 Wörtern. Die Aufgaben wurden danach in Kooperation mit Dr. Ulf Kröhne vom TBA-Forschungszentrum am DIPF in eine computerbasierte Testumgebung (CBA Item-Builder) eingebettet (AP3).

Auf Basis von existenten Forschungsergebnissen zu Qualitätsdimensionen bei integrierten Aufgaben (Plakans, 2009; Plakans & Gebril, 2013, 2017; Gebril & Plakans, 2014) wurde parallel zur Entwicklung der Aufgaben ein erstes Bewertungsschema erstellt (*AP1*). Konkret wurden zu den drei angenommenen Qualitätsdimensionen (a) Qualität der Ausgangstextnutzung, (b) Diskurssynthese und (c) Sprachqualität jeweils mehrere Kriterien mit Deskriptoren für eine fünfstufige Bewertung entwickelt: (1) Identifikation relevanter Ideen des Ausgangstextes, (2) Korrektheit und Präzision der Ideen des Ausgangstextes, (3) sprachliche Verarbeitung (Paraphrasieren vs. wörtliche Übernahme), (4) Zitieren aus dem Ausgangstext (nur Stellungnahme), (5) Synthese von Ideen des Ausgangstextes und eigenen Ideen (nur Stellungnahme), (6) Textstruktur, thematische Entwicklung und Kohärenz, (7) Kohäsion, (8) Wortschatz und (9) Grammatik.

Danach wurden das Aufgabenmaterial und das Bewertungsschema in einem mehrstufigen Prozess erprobt und überarbeitet. Unter anderem fand ein Expert*innen-Workshop (AP5) statt und es gab eine Pilotierung mit teilnehmenden Studierenden (*AP4*), bei der Lautes-Denken-Protokolle und Fragebögen durchgeführt wurden. Anschließend wurden anhand der Schreibprodukte aus der Pilotierung zehn Bewerter*innen geschult (*AP5*), wobei das Bewertungsschema auf Basis der Ratingdaten und Rückmeldungen der Bewerter*innen nochmals revidiert wurde. Schließlich erfolgte anhand von mehrfach bewerteten Texten eine erste Einschätzung der Übereinstimmung der Bewertungen und damit eine Auswahl der Bewerter*innen für die Hauptuntersuchung. Zur Entwicklung der Bewertungsskala wurde in Harsch et al. (2024a) publiziert, die Validierung erfolgte in Kooperation mit Dr. Voula Kanistra und wurde in Harsch et al. (2024b) publiziert.

Auf dieser Grundlage wurde die Hauptuntersuchung durchgeführt, in der die entwickelten integrierten Aufgaben mit unabhängigen Lese- und Schreibtests des *Integrated Skills of Englisch Tests* des Trinity College London (TCL) ergänzt wurden (*AP2*), welche ebenfalls in die Online-Testumgebung eingepflegt wurden.Jede*r Teilnehmer*in hatte drei Aufgaben zu bearbeiten (*AP8*). Insgesamt nahmen 451 Personen teil, darunter 127 Schüler*innen in Englisch-Leistungskursen und 324 Studierende im ersten Semester englischsprachiger Bachelorprogramme. Insgesamt erhielten wir 674 integrierte Schreibprodukte, 269 Antworten zu den Lesetest und 252 isolierte Schreibprodukte. Die isolierten Lese- und Schreibtests wurden direkt von TCL ohne Mehrkosten für uns ausgewertet. Die integrierten Schreibprodukte wurden von den o.e. sechs geschulten Bewerter*innen beurteilt (*AP9*). Um die Übereinstimmung und Konsistenz der Bewertungen zu prüfen, wurden 72 Schreibprodukte sowohl zu Beginn als auch am Ende des Bewertungsverfahrens von allen Bewerter*innen beurteilt, während die übrigen Texte jeweils einzeln bewertet wurden. Die Beurteilerübereinstimmungen erwiesen sich bei einigen Kriterien als unzureichend. Konkret war die Reliabilität der Gesamturteile der sechs Bewerter*innen mit ICCs von > .80 überwiegend hoch, während die Werte für die Einzelurteile mit ICCs von < .60 fast durchgehend im unbefriedigenden Bereich lagen.

Die *Konstrukt-Dimensionalität* wurde auf Basis der Auswertungen der Hauptuntersuchung empirisch überprüft (*AP10*). Konfirmatorische Faktorenanalysen zeigten, dass die Qualitätskriterien nicht der ursprünglich angenommenen dreidimensionalen Struktur entsprechen (Delgado-Osorio et al., 2022b). Für beide Aufgabentypen (Zusammenfassung und Stellungnahme) lassen sich stattdessen jeweils zwei Dimensionen identifizieren, wobei sich die Strukturen im Detail unterscheiden. Zudem konnte festgestellt werden, dass die Zuordnung einzelner Kriterien zu diesen Dimensionen auf dieser Datenbasis nicht eindeutig ausfällt.

Darüber hinaus wurde der *Zusammenhang der erfassten Kompetenzen mit unabhängigen Lese- und Schreibfähigkeiten sowie mit Indikatoren des Studienerfolgs untersucht*. Wie erwartet lassen sich die Leistungen in den integrierten Aufgaben zu einem signifikanten, aber nur begrenzten Umfang (alle $R^2 < .20$) durch die Leistungen in den isolierten Lese- und Schreibtests vorhersagen (Harsch et al., 2022). Außerdem nahmen 70 Personen an einer Nachbefragung (AP11) teil, die sechs bis neun Monate nach Ende der Hauptuntersuchung durchgeführt wurde. Dabei wurden Selbsteinschätzungen zu Leistungen im Studium sowie zur Bewältigung von Anforderungen im englischsprachigen Studium erhoben. Die Analysen dieser Daten (Harsch et al., 2022; Hartig et al., 2022) zeigen nur geringe Zusammenhänge der Studienerfolgsvariablen mit der Qualität der Leistungen in den integrierten Aufgaben.

In **MASK II** wurde im Hinblick auf Kompetenzmodell und Diagnostik eine verbesserte Datengrundlage der Bewertungen der integrierten Schreibprodukte geschaffen. Damit wurde sowohl eine vertiefende Untersuchung der Qualität menschlicher Urteile von Bewerter*innen mit und ohne Expertise als auch weitere Analyse der Konstrukt-Dimensionalität des integrierten Schreibens ermöglicht.

Zu diesem Zweck rekrutierten und schulten wir vier erfahrene Bewerter*innen (*AP2*). Konkret luden wir Englischlehrer*innen von Hochschulen und Sprachenzentren in Deutschland ein, die sich auf Englisch für akademische Zwecke spezialisiert haben und Erfahrung mit der Bewertung integrierter Lese-/Schreibaufgaben besitzen. Nach einem Auswahlgespräch nahmen die vier ausgewählten Lehrkräfte an einer Rater-Schulung teil, um sich mit unseren Bewertungsinstrumenten und dem Herangehen vertraut zu machen; die Schulung verlief weitgehend parallel zu der Schulung in MASK I. Anschließend beurteilten die Lehrkräfte alle 674 integrierten Schreibprodukte der MASK I-Hauptuntersuchung (*AP3*). Da wir nun mehrere Bewertungen pro Text hatten, führten wir erneut Beurteiler-Übereinstimmungsanalysen durch, um die Zuverlässigkeit der Bewertungen zu prüfen. Basierend auf den Mehrfachbewertungen, einschließlich der Bewertungen aus MASK I durch die sechs Novizen-Bewerter*innen sowie der Bewertungen durch die vier erfahrenen Bewerter*innen, erhielten wir für alle Kriterien gute bis ausgezeichnete ICC (*2,k)* Werte für mehrere Bewerter*innen im Bereich von 0.78 bis 0.96.

Auf Basis dieser soliden Datengrundlage haben wir eine vertiefende Untersuchung der Dimensionalität menschlicher Urteile (*AP7*) durchgeführt. Die mit den Ratings aus der ersten Projektphase durchgeführten konfirmatorischen Faktorenanalysen wurden wiederholt, auch hier konnte die dreidimensionale Struktur nicht in Gänze bestätigt werden. Für die Zusammenfassung ergab sich eine zweidimensionale Struktur, bei der Diskurssynthese und Sprachqualität zu einem Faktor zusammenfallen, für die Stellungnahme wurden die drei Dimensionen beibehalten. Für beide Modelle mussten mehrere Residualkovarianzen zwischen einzelnen Indikatoren freigesetzt werden, um eine akzeptable Modellpassung zu erreichen. Auf Basis der modifizierten Modelle wurde auch der Zusammenhang der Leistungen in den integrierten Aufgaben mit den unabhängigen Lese- und Schreibfähigkeiten erneut analysiert, hierbei wurden auch die unabhängigen sprachlichen Fähigkeiten als latente Variablen modelliert. Es zeigte sich, dass bei gleichzeitiger Berücksichtigung nur die Lesefähigkeit, nicht aber die Schreibfähigkeit prädiktiv für die Leistungen in den integrierten Aufgaben ist. Die mit Abstand größte Varianzaufklärung fand sich für die Qualität der Ausgangstextnutzung bei Zusammenfassungen (latentes $R^2 = .65$), die niedrigste für Diskurssynthese bei Stellungnahmen (latentes $R^2 = .31$). Die Ergebnisse der Analysen wurden von Harsch et al. (2026) präsentiert. Das Manuskript zur Publikation dieser Ergebnisse ist zum Zeitpunkt der Berichtlegung in Arbeit und soll in *Language Testing* oder *Language Assessing Quarterly* eingereicht werden.

b) Bewertungsprozesse und Expertise-Effekte der Bewerter*innen: Untersuchung der menschlichen Urteile und Bewertungsprozesse von Bewerter*innen mit und ohne Expertise.

Im Rahmen von MASK I und MASK II wurden die Bewertungsprozesse der Bewerter*innen und deren Wahrnehmungen mithilfe von Lautes-Denken-Protokollen und halbstrukturierten Fokusgruppen-Interviews (FGIs) untersucht. Die Protokolle folgten in beiden Projektphasen demselben Leitfaden und nutzten dieselben Instrumente. Unabhängig von der Projektphase bewerteten die Bewerter*innen in den individuellen Lautes-Denken-Sitzungen dieselben vier Texte und verbalisierten dabei ihre Bewertungsprozesse. Die Texte unterschieden sich hinsichtlich der Aufgabentypen und des Ausmaßes der Textübernahmen aus dem Ausgangstext. In den halbstrukturierten FGIs hatten die Rater*innen die Möglichkeit, Aspekte der Bewertungsskala, Aufgabenmerkmale, Machbarkeit sowie die verwendeten Bewertungsstrategien zu diskutieren. Die Datenerhebung fand online über Zoom (Zoom Video Communications Inc., 2020) statt. Alle Sitzungen wurden aufgezeichnet, transkribiert, anonymisiert und anschließend in MAXQDA (VERBI Software, 2020) ausgewertet.

In **MASK I** wurden während der Pilotierungsphase erste Lautes-Denken-Protokolle eingesetzt, um die Entwicklung bzw. Anpassung der Bewertungsskala zu unterstützen. Nach der Bewertung in der Hauptstudie wurden weitere Lautes-Denken-Protokolle sowie Fokusgruppen-Interviews durchgeführt, um Zusammenhänge zwischen Bewertungsverhalten und

Bewertungsqualität zu untersuchen und die Validierung der Bewertungsskala zu unterstützen (Harsch et al., 2024b). Die Lautes-Denken-Protokolle und Fokusgruppen-Interviews aus MASK I mit fünf Bewerter*innen bilden die Grundlage für die zentralen Ergebnisse des Projekts hinsichtlich der Validität der Bewertungsskala sowie des Wesens des Bewertungsprozesses. Wie Harsch et al. (2024b) berichten, empfanden die Bewerter*innen das Bewertungssystem als geeignet zur Beurteilung integrierter Aufgaben. Sie verfügten über ein klares Verständnis dafür, auf welche Merkmale in den Leistungen zu fokussieren sind, und konnten zwischen den Kriterien differenzieren. Gleichzeitig zeigten sich Unsicherheiten bei der praktischen Anwendung der Bewertungsskala, insbesondere bei der Beurteilung des Kriteriums „Ausgangstextnutzung“. Die Bewerter*innen betonten jedoch, dass diese Herausforderungen weniger auf die Skala selbst als vielmehr auf die inhärente Komplexität integrierter Aufgaben zurückzuführen seien.

Ein zentrales Ziel von **MASK II** war der quantitative und qualitative Vergleich von Bewerter*innen mit und ohne Expertise. Hierfür wurden weitere Lautes-Denken-Protokolle sowie Fokusgruppen-Interviews mit vier erfahrenen Bewertern*innen durchgeführt (*AP4*). Ergänzend wurden 674 Bewertungen von Bewerter*innen mit unterschiedlichem Erfahrungsstand einbezogen. Dies ermöglichte eine Mixed-Methods-Studie (Koval et al., in Revision), in der die Daten (Lautes-Denken-Protokolle, Fokusgruppen-Interviews und Bewertungen von Novizen-Bewerter*innen) aus MASK I mit denen aus MASK II verglichen wurden. Der quantitative Teil der Studie untersuchte Unterschiede zwischen den Gruppen hinsichtlich der Interrater-Übereinstimmung, der Bewertungsstrenge und der Halo-Effekte (*AP6*). Für den qualitativen Teil wurde – auf Grundlage früherer Forschung (z. B. Crisp, 2008; Zhang, 2016; Wolfe et al., 1998) – eine Reihe qualitativer Variablen identifiziert, die zur Erklärung der quantitativen Ergebnisse beitragen könnten (*AP5*). Die Ergebnisse zeigen, dass Expertise allein keine höhere Bewertungsqualität voraussetzt. Hinsichtlich der Übereinstimmung und der Strenge traten größere Unterschiede zwischen einzelnen Bewerter*innen als zwischen den Expert*innen- und Nicht-Expert*innen-Gruppen auf. Bestimmte Bewertungsverhaltensweisen, etwa das Ausmaß des Lesens der Studententexte oder das Evaluieren spezifischer Merkmale, spielten dabei eine zentrale Rolle. Beim Halo-Effekt hingegen ließen sich Gruppentendenzen erkennen: Expert*innen zeigten häufiger Halo-Effekte. Dies könnte darauf zurückzuführen sein, dass sie, laut eigenen Reflexionen, stärker mit holistischen Ansätzen vertraut waren. Insgesamt deuten die Ergebnisse darauf hin, dass auch unerfahrene Bewerter*innen für analytische Bewertungen integrierter Schreibaufgaben geeignet sein können. Mit angemessener Schulung entwickeln sie ein nachvollziehbares Bewertungsverhalten sowie ein metakognitives Bewusstsein und erreichen eine Bewertungsqualität, die der erfahrener Expert*innen entspricht oder diese sogar übertrifft.

Um zur Wiederverwendbarkeit und offenen Zugänglichkeit von Forschungsdaten beizutragen, sind wir derzeit dabei, die Textdaten von Schüler*innen und Studierende und die dazugehörigen Bewertungen im FD-LEX am Mercator-Institut für Sprachwissenschaften über den Verbund Forschungsdaten Bildung (VerbundFDB) zu veröffentlichen.

c) Bearbeitungsprozesse der Studierenden: Untersuchung der Prozesse von Studierenden bei der Aufgabenbearbeitung.

Im Hinblick auf die Untersuchung der Bearbeitungsprozesse der Studierenden wurden in **MASK I** die Lautes-Denken-Protokolle aus der Pilotierung (*AP4*) analysiert. Diese wurden online über eine Videokonferenzplattform durchgeführt, die es erlaubte, den Bildschirm der Teilnehmenden während der Aufgabenbearbeitung aufzuzeichnen (Zoom Video Communications Inc., 2020). Während die Studierenden ihre Gedanken und Vorgehensweisen laut äußerten, wurden parallel audiovisuelle Daten in Form von Sprach- und Bildschirmaufzeichnungen erfasst. Insgesamt wurden 14 Protokolle transkribiert und anschließend sowohl deduktiv als auch induktiv mit Hilfe der thematischen Analyse (Braun &

Clarke, 2006) in MAXQDA ausgewertet. Ziel war es, Bearbeitungsprozesse sichtbar zu machen, die im finalen Schreibprodukt nicht erkennbar sind, und damit zu einem vertieften Verständnis des zugrunde liegenden Konstrukts beizutragen. Die qualitativen Analysen (*AP6*) zeigten die Bedeutung metakognitiver und lesebezogener Strategien, die je nach Aufgabentyp, Zusammenfassung oder Stellungnahme, unterschiedlich zum Einsatz kamen. Die Ergebnisse dieser Analysen wurden in Delgado-Osorio et al. (2022a) und Delgado-Osorio et al. (2023a) publiziert.

Darüber hinaus wurden im Rahmen der Hauptuntersuchung (*AP8*) zur Bearbeitung integrierter Aufgaben durch die Studierenden verschiedene Verhaltensweisen (Logdaten) erfasst, darunter die Verwendung von Annotationsstrategien (Markierungen und Notizen). Um zu untersuchen, welche Strategien die Studierenden bei der Bearbeitung integrierter Aufgaben anwenden und wie effektiv diese sind, wurde spezifisch die Verwendung von Markierungen und Notizen analysiert, sowie die Zusammenhänge zwischen den annotierten Inhalten und der Qualität der Schreibprodukte (Delgado-Osorio et al., 2023b, 2025d). Im Gegensatz zu qualitativen Daten erlauben die Logdaten die Analyse von Bearbeitungsprozessen auf der gesamten Stichprobe (674 integrierte Schreibprodukte).

In **MASK II** wurden weitere Analysen der Logdaten durchgeführt (*AP1*). Ergänzend zu MASK I haben wir die Analyse der Logdaten auf andere Verhaltensweisen erweitert (z. B. Keystrokes, Klicks auf Buttons, Wörterbuchbenutzung und Scrolling). Die Analysen fokussierten darauf, inwieweit Bearbeitungsmuster in den Logdaten identifiziert werden konnten und welchen Zusammenhang es zwischen diesen Mustern und der Qualität der Produkttexte gibt. Zu diesem Zweck haben wir eine Studie mit Sequenzanalyse durchgeführt, deren Ergebnisse in Delgado-Osorio et al. (2025e) veröffentlicht wurden. Die Ergebnisse zeigten vier Bearbeitungsmuster in unseren integrierten Aufgaben, die sich in der Häufigkeit und dem Zeitpunkt bestimmter Verhaltensweisen unterscheiden und einen Zusammenhang mit der Qualität der verfassten Texte aufweisen. Um eine transparente Nachvollziehbarkeit sowie eine Nachnutzung der angewandten Methode zu ermöglichen, ist das vollständig dokumentierte R-Skript der Sequenzanalyse Open Access auf Zenodo verfügbar (Delgado-Osorio & Hartig, 2025).

d) Computergestützte & automatisierte Textbewertung: Entwicklung und Prüfung computergestützter Verfahren zur Bewertungsunterstützung und Automatisierung der Textbewertung.

Im Hinblick auf die computergestützte und automatisierte Textbewertung lag der Schwerpunkt von **MASK I** darauf, die computergestützte Unterstützung menschlicher Bewerter*innen bei der Beurteilung der Verwendung von Quelltexten zu untersuchen und einen ersten Ansatz zur Automatisierung der Textbewertung zu erforschen.

Zur Unterstützung menschlicher Bewerter*innen wurde in Zusammenarbeit mit Torsten Zesch ein Software-Tool entwickelt (*AP7*), das Passagen in den Schreibprodukten markiert, die als wortwörtliche Textübernahme identifiziert wurden. Die Auswirkungen des Tools auf die Qualität der Bewertungen wurde während der der Ratings der Textprodukte aus der Hauptuntersuchung (*AP9*) und anhand der quantitativen Analyse der Rating-Daten aus der Hauptuntersuchung (*AP10*) untersucht. Zu diesem Zweck bewerteten die Bewerter*innen 72 Schreibprodukte mit dem Tool und 94 ohne das Tool, was zu folgenden Ergebnissen führte: Koval et al. (2022) zeigten, dass die Interrater-Reliabilität beim Kriterium der sprachlichen Verarbeitung (d.h., der Art und des Umfangs von Paraphrasen statt wörtlicher Übernahme), durch die Unterstützung des Tools verbessert wurde. Besonders bei der Bewertung von Zusammenfassungsaufgaben scheint die Qualität der Bewertungen durch den Einsatz des Tools stärker zu profitieren, was auf die spezifische Beschaffenheit dieser Aufgaben zurückzuführen ist.

Parallel dazu wurde ein erster Ansatz zur automatisierten Textanalyse an den Schreibprodukten getestet (*AP7*). Zu diesem Zweck wurden Indizes zur Verarbeitung natürlicher Sprache (NLP) mithilfe der Softwares TAACO, TAALES und SCA sowie T-MITOCAR in Zusammenarbeit mit Pablo Pirnay-Dummer ermittelt. Diese Tools liefern eine Reihe von NLP-Indizes zur semantischen Nähe zwischen Ausgangstexten und Schreibprodukten sowie Indizes zur Kohäsion und Sprachqualität. Laut bisheriger Literatur könnten diese Indizes mit den Kriterien unseres Bewertungsschemas in Zusammenhang stehen. Zusätzlich erhielten wir einen Prozentsatz der wortwörtlichen Übernahmen aus der in Zusammenarbeit mit Torsten Zesch entwickelten o.g. Software. Analysen der Zusammenhänge zwischen automatisierten Textbewertungsindizes und menschlichen Beurteilungen wurden mithilfe von Korrelations- und Regressionsanalysen durchgeführt (Hartig, 2022). Die Ergebnisse zeigten, dass für die meisten Kriterien zwischen 20% und 50% der Varianz durch Modelltraining und Vorhersagen auf verschiedenen Teilmengen von Daten erklärt werden konnten. Der Prozentsatz der wörtlichen Übernahmen erwies sich dabei als wichtiger Prädiktor. Darüber hinaus stellten wir fest, dass Zusammenfassungen viel genauer mit automatisierten NLP-Indizes bewertet werden können als Stellungnahmen.

Ursprünglich war im Rahmen von **MASK II** geplant, die Untersuchung, wie NLP-basierte Textmerkmale menschliche Bewertungen vorhersagen können (*AP8*), mithilfe komplexerer Vorhersagemodelle des maschinellen Lernens in Kooperation mit Pablo Pirnay-Dummer auszuweiten. Aufgrund der rasanten Entwicklung von KI-Methoden als fortgeschrittene Anwendung von NLP kollaborierten wir jedoch mit Sebastian Gombert (DIPF), um KI-Modelle (BERT-basierte Modelle) zur Vorhersage unserer menschlichen Bewertungen zu untersuchen. Insgesamt zeigten die Analysen, dass sich die Bewertungen integrierter Schreibaufgaben mit automatisierten Bewertungsmodellen gut vorhersagen lassen. Am besten funktionieren Modelle, bei denen alle Aufgaben in einem einzigen Modell integriert sind und die die Beziehung zwischen Text und Bewertungskriterien explizit darstellen. Die Vorhersage-Genauigkeit variierte jedoch je nach Kriterium: Die höchsten Korrelationen zwischen den Modellvorhersagen und den menschlichen Bewertungen wurden bei den Kriterien „Zitieren aus dem Ausgangstext“ (r = 0.89), „Grammatik” (r = 0.81) und „Wortschatz” (r = 0.81) festgestellt, die niedrigsten bei den Kriterien „Synthese von Ideen des Ausgangstextes und eigenen Ideen“ (r = 0.63) und „Korrektheit und Präzision der Ideen des Ausgangstextes“ (r = 0.64). Das Manuskript zur Veröffentlichung dieser Ergebnisse befindet sich zum Zeitpunkt der Berichterstattung in Arbeit und wird bei *Assessing Writing* eingereicht (Delgado-Osorio, Gombert et al., 2026).

Anwendungsperspektiven

Das im Projekt entwickelte Tool zur Erkennung von wortwörtlich übernommenen Passagen aus Quellentexten wird im Sprachzentrum der Hochschulen im Land Bremen im Schreibunterricht eingesetzt. Wir arbeiten derzeit daran, das auf unserem Server gehostete Tool auch anderen Sprachenzentren zugänglich zu machen.
Die Erkenntnisse zur Nutzung von KI-Modellen zur Bewertung integrierter Schreibleistungen können genutzt werden, um einen Algorythmus zu trainieren, der automatisiertes Feedback geben kann.

Literaturverzeichnis

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, *38*(2), 247–264. https://doi.org/10.1080/03057640802063486

Delgado-Osorio, X., & Hartig, J. (2025). *Students' behavioral patterns in integrated writing tasks: A sequence analysis approach* [Datenanalyse]. Zenodo. https://doi.org/10.5281/zenodo.14916920

Delgado-Osorio, X., Gombert, S., Hartig, J., Drachsler, H., & Harsch, C. (2026). *Automated Rubric-based Analytic Scoring of Summative and Argumentative Writing: Accuracy and Multi-Task Transferability of Pre-trained Transformer Encoder Models When Predicting Human Ratings* [Manuscript in preparation].

Delgado-Osorio, X., Koval, V., Hartig, J., & Harsch, C. (2022a, 10.–15. September). Source-Text Strategic Processing in Reading-into-Writing Tasks: A Comparison between Summary and Argumentative Tasks [Konferenzbeitrag]. *DGPs-Kongress: view on|of science*, Hildesheim, Deutschland. https://www.uni-hildesheim.de/dgps2022/

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2022b, 9.–11. März). Dimensionen der Textqualität von integrierten Schreibprodukten in Englisch als Fremdsprache [Konferenzbeitrag]. *9. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF)*: *Alles auf Anfang? Bildung im digitalen Wandel*, online. https://gebf2022.de/

Delgado-Osorio, X., Koval, V., Hartig, J., & Harsch, C. (2023a). Strategic processing of source text in reading-into-writing tasks: A comparison between summary and argumentative tasks. *Journal of English for Academic Purposes*, 62, 101227. https://doi.org/10.1016/j.jeap.2023.101227

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2023b, 28. Februar–2. März). Die Rolle von Lesestrategien in integrierten Schreibaufgaben für Englisch als Fremdsprache [Konferenzbeitrag]. *10. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF),* Essen, Deutschland. https://www.uni-due.de/gebf_2023/index

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2025e). Students' behavioral patterns in integrated writing tasks: A sequence analysis approach. *Journal of Educational Psycholo-gy*, 117(6), 898–917. https://doi.org/10.1037/edu0000957

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. https://doi.org/10.1016/j.asw.2014.03.002

Harsch, C., Hartig, J., Koval, V., & Delgado-Osorio, X. (2022, 7.–11. März). Linguistic preparedness in EMI contexts [Konferenzbeitrag]. *43rd Language Testing Research Colloquium (LTRC)*, online. https://www.iltaonline.com/page/LTRC2022

Harsch, C., Koval, V., Delgado-Osorio, X., & Hartig, J. (2024a). Usability of CEFR Companion Volume scales for the development of an analytic rating scale for academic integrated writing assessment. *CEFR Journal – Research and Practice*, 6, 155–177. https://doi.org/10.37546/JALTSIG.CEFR6-9

Harsch, C., Koval, V., Kanistra, P., & Delgado-Osorio, X. (2024b). Validating an integrated reading-into-writing scale with trained university students. *Assessing Writing*, 62, 100894. https://doi.org/10.1016/j.asw.2024.100894

Hartig, J. (2022, April). Struktur von menschlichen Urteilen und automatisierten Indices zur Bewertung der Textqualität bei integrierten Aufgaben [Vortrag]. *Forschungskolloquium Methoden der empirischen Bildungsforschung*, Hirschegg, Österreich.

Hartig, J., Delgado-Osorio, X., Harsch, C., & Koval, V. (2022, September 10-15). Erfassung von akademisch-schriftsprachlichen Kompetenzen im Englischen mit integrierten Schreibaufgaben [Konferenzbeitrag]. *Kongress der Deutschen Gesellschaft für Psychologie*, Hildesheim, Deutschland.

Koval, V., Delgado-Osorio, X., Harsch, C., & Hartig, J. (2022, 3.–5. Juni). Adjustments of rating practices to improve assessment quality of academic integrated writing [Konferenzbeitrag]. *18th European Association for Language Testing and Assessment Conference (EALTA): Evolving contexts: flexibility and adaptation in language testing and assessment,* Budapest, Ungarn. https://ealta.eu/conferences/18th-ealta-conference-3-5-june-2022-budapest-evolving-contexts-flexibility-and-adaptation-in-language-testing-and-assessment/

Koval, V., Delgado-Osorio, X., Harsch, C., Hartig, J., & Kanistra, Paraskevi. (2025). *Beyond Experience: Rating Quality and Rater Behaviors in Integrated Writing* [Manuscript in revision].

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. https://doi.org/10.1177/0265532209340192

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing*, *31*, 98–112. https://doi.org/10.1016/j.asw.2016.08.005

VERBI Software. (2020). *MAXQDA 2020* [Computer software]. https://www.maxqda.com

Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465–492. https://doi.org/10.1177/0741088398015004002

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37–53. https://doi.org/10.1016/j.asw.2015.11.001

Zoom Video Communications Inc. (2020) [Computer software]. https://zoom.us/

4 Veröffentlichte Projektergebnisse

4.1 Kategorie A – Fachaufsätze in Peer Review-Zeitschriften, Beiträge zu Konferenzen mit Peer Review oder Sammelbänden sowie Buchpublikationen

Peer Review-Zeitschriften

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2025a). Students' behavioral patterns in integrated writing tasks: A sequence analysis approach. *Journal of Educational Psychology*, 117(6), 898–917. https://doi.org/10.1037/edu0000957

[Open-Access] Harsch, C., Koval, V., Delgado-Osorio, X., & Hartig, J. (2024a). Usability of CEFR Companion Volume scales for the development of an analytic rating scale for academic integrated writing assessment. *CEFR Journal – Research and Practice*, 6, 155–177. https://doi.org/10.37546/JALTSIG.CEFR6-9

[Open-Access] Harsch, C., Koval, V., Kanistra, P., & Delgado-Osorio, X. (2024b). Validating an integrated rea-ding-into-writing scale with trained university students. *Assessing Writing*, 62, 100894. https://doi.org/10.1016/j.asw.2024.100894

Delgado-Osorio, X., Koval, V., Hartig, J., & Harsch, C. (2023a). Strategic processing of source text in reading-into-writing tasks: A comparison between summary and argumentative tasks.

Journal of English for Academic Purposes, 62, 101227. https://doi.org/10.1016/j.jeap.2023.101227

Beiträge zu Konferenzen mit Peer Review

Delgado-Osorio, X., Koval, V., Hartig, J., & Harsch, C. (2025b, 25.–29. August). Novice vs. Expert Raters Performance When Evaluating Integrated Writing Task: A Mixed-Methods Study [Konferenzbeitrag]. *European Association for Research on Learning and Instruction (EARLI) Conference*, Graz, Österreich. https://www.earli.org/events/earli2025

Koval, V., Delgado-Osorio, X., Harsch, C., & Hartig, J. (2025, 26.–31. Mai). Novice vs. expert raters' performance when evaluating integrated writing tasks: A mixed-methods study [Konferenzbeitrag]. *21st European Association for Language Testing and Assessment Conference (EALTA): Making an impact? Teaching, learning, and assessment*, Salzburg, Österreich. https://ealta-annual-conference.org/

Harsch, C. (2025, 4.- 5 März) Integrated reading-into-writing – how to assess the complexities? [Impulsvortrag] *New Directions Conference 2025*, British Council, Berlin, Germany. https://www.britishcouncil.de/en/events/new-directions-conference-2025-berlin

Delgado-Osorio, X., Hartig, J., Koval, V., & Harsch, C. (2025c, 27.–29. Januar). Effekte von professioneller Expertise auf die Konsistenz und Strenge bei der Bewertung integrierter Schreibprodukte [Konferenzbeitrag]. *12. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF)*, Mannheim, Deutschland. https://www.gebf2025.de

Delgado-Osorio, X., Koval, V., Hartig, J., & Harsch, C. (2024a, 1.–5. Juli). Sequence analysis of log data: an application example from a study of integrated writing [Konferenzbeitrag]. *45th Language Testing Research Colloquium (LTRC)*, Innsbruck, Österreich. https://www.uibk.ac.at/en/congress/ltrc2024/

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2024b, 18.–20. März). Eine Sequenzanalyse von Prozessen und Strategien bei der Bearbeitung integrierter Schreibaufgaben im English for Academic Purposes [Konferenzbeitrag]. *11. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF)*, Potsdam, Deutschland. https://www.uni-potsdam.de/de/gebf2024/index

Harsch, C., Koval, V., Hartig, J., & Delgado-Osorio, X. (2023, 5.–9. Juni). Validation of a rating scale for computer- based integrated assessment in a lingua franca context [Konferenzbeitrag]. *44th Language Testing Research Colloquium (LTRC)*, New York, Vereinigte Staaten. https://ltrc2023.weebly.com/

Koval, V., Delgado-Osorio, X., Harsch, C., & Hartig, J. (2023, 5.–9. Juni). Online implementation of a project for an integrated writing assessment [Konferenzbeitrag]. *44th Language Testing Research Colloquium (LTRC)*, New York, Vereinigte Staaten. https://ltrc2023.weebly.com/

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2023b, 28. Februar–2. März). Die Rolle von Lesestrategien in integrierten Schreibaufgaben für Englisch als Fremdsprache [Konferenzbeitrag]. *10. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF),* Essen, Deutschland. https://www.uni-due.de/gebf_2023/index

Koval, V., Delgado-Osorio, X., Harsch, C., & Hartig, J. (2023). Online implementation of an integrated writing assessment project [Konferenzbeitrag]. *8th Bremen Symposium*, Bremen, Germany

Delgado-Osorio, X., Koval, V., Hartig, J., & Harsch, C. (2022, 10.–15. September). Source-Text Strategic Processing in Reading-into-Writing Tasks: A Comparison between Summary and Argumentative Tasks [Konferenzbeitrag]. *DGPs-Kongress: view on|of science*, Hildesheim, Deutschland. https://www.uni-hildesheim.de/dgps2022/

Delgado-Osorio, X., Hartig, J., Harsch, C., & Koval, V. (2022, 9.–11. März). Dimensionen der Textqualität von integrierten Schreibprodukten in Englisch als Fremdsprache [Konferenzbeitrag]. *9. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF)*: *Alles auf Anfang? Bildung im digitalen Wandel*, online. https://gebf2022.de/

Koval, V., Delgado-Osorio, X., Harsch, C., & Hartig, J. (2022, 3.–5. Juni). Adjustments of rating practices to improve assessment quality of academic integrated writing [Konferenzbeitrag]. *18th European Association for Language Testing and Assessment Conference (EALTA): Evolving contexts: flexibility and adaptation in language testing and assessment,* Budapest, Ungarn. https://ealta.eu/conferences/18th-ealta-conference-3-5-june-2022-budapest-evolving-contexts-flexibility-and-adaptation-in-language-testing-and-assessment/

Hartig, J. (2022, April). Struktur von menschlichen Urteilen und automatisierten Indices zur Bewertung der Textqualität bei integrierten Aufgaben [Vortrag]. *Forschungskolloquium Methoden der empirischen Bildungsforschung*, Hirschegg, Österreich.

Hartig, J., Delgado-Osorio, X., Harsch, C., & Koval, V. (2022, September 10-15). Erfassung von akademisch-schriftsprachlichen Kompetenzen im Englischen mit integrierten Schreibaufgaben [Konferenzbeitrag]. *Kongress der Deutschen Gesellschaft für Psychologie*, Hildesheim, Deutschland.

Harsch, C., Hartig, J., Koval, V., & Delgado-Osorio, X. (2022, 7.–11. März). Linguistic preparedness in EMI contexts [Konferenzbeitrag]. *43rd Language Testing Research Colloquium (LTRC)*, online. https://www.iltaonline.com/page/LTRC2022

Delgado-Osorio, X. (2021, 22.–24. September). Strategic Processing in an Integrated Writing Task for Academic English [Konferenzbeitrag]. *DGFF-Kongress*, online. https://kongress.dgff.de/

Harsch, C., Hartig, J., Koval, V., & Delgado-Osorio, X. (2021, 4.–5. Juni). Variations of common expectations – linguistic preparedness for EMI contexts [Konferenzbeitrag]. *17th European Association for Language Testing and Assessment Conference (EALTA): Variations of common standards,* online. https://ealta.eu/conferences/17th-ealta-conference-variations-on-common-standards-4-5-june-2021-online-conference/

4.2 Kategorie B – Jede weitere Form öffentlich gemachter Ergebnisse

Delgado Osorio, X. (2024). *Strategic processing of integrated writing tasks in the context of English for academic purposes* [Dissertation]. Johann Wolfgang Goethe-Universität, Frankfurt am Main. https://doi.org/10.1016/j.jeap.2023.101227

Delgado-Osorio, X., & Hartig, J. (2025). *Students' behavioral patterns in integrated writing tasks: A sequence analysis approach* [Datenanalyse]. Zenodo. https://doi.org/10.5281/zenodo.14916920

Delgado-Osorio, X., Hartig, J. Koval, V. & Harsch, C. (2025d) The use of highlighting and note-taking by EMI students in reading-into-writing tasks. *Language Learning in Higher Education*. Das eingereichte Manuskript ist dem Bericht beigefügt.

Harsch, C., Hartig, J., Delgado-Osorio, X. & Koval, V. (2026). *Dimensionalitäten integrierter Lese-Schreibkompetenzen - Braucht es integrierte Tests überhaupt?* Vortrag auf dem Kolloquium „Methoden der empirischen Bildungsforschung“, 18. bis 23. Januar 2026 in Riezlern, Österreich.

Koval, V., Delgado-Osorio, X., Harsch, C., Hartig, J., & Kanistra, Paraskevi. (2025). *Beyond Experience: Rating Quality and Rater Behaviors in Integrated Writing* [Manuscript in revision]. Das eingereichte Manuskript ist dem Bericht beigefügt.

Koval, V., Delgado-Osorio, X., Harsch, C., & Hartig, J. (2024). Online implementation of an integrated writing assessment project. *8th Bremen Symposium 2023 Conference Proceedings*.

Koval, V. (2023). Exploring anti-plagiarism tool effects in the assessment of academic reading-into-writing. *ALTE 2023 Conference Proceedings*.

4.3 Patente (angemeldete und erteilte) [entfällt]

<NICHT ÖFFENTLICHER TEIL>

5 Weitere Informationen zu Projekt, Qualifikationen und Ausblick

Aufgrund der Covid-19-Pandemie wurde die Datenerhebung anders als geplant von Anfang an – auch bei der Aufgabenerprobung und Pilotierung – internetbasiert durchgeführt. Dies erforderte eine Umsetzung der Tests in eine computerbasierte Umgebung schon in der Pilotierungsphase. Ursprünglich war eine computerbasierte Testung erst für die Hauptstudie vorgesehen. Die Entwicklung der Testumgebung erfolgte mit Unterstützung von Kolleg*innen des TBA-Zentrums am DIPF im CBA ItemBuilder. Die browserbasierte Bearbeitung der Testaufgaben erlaubte den Proband*innen unter anderem das Markieren in den Ausgangstexten und das Einfügen von Anmerkungen und Notizen.

Auch hinsichtlich der Stichprobenrekrutierung ergaben sich durch die Pandemie Änderungen. Während ursprünglich Datenerhebungen an Hochschulen und Gymnasien vor Ort an den beiden Standorten geplant waren, wurde nun die gesamte Datenerhebung online durchgeführt. Die Rekrutierung erfolgte über Sprachenzentren von Hochschulen und über Schul-Lehrkräfte von Englisch-Leistungskursen. Insgesamt konnten in einem Zeitraum von vier Monaten (nach Datenbereinigung) N = 451 Teilnehmende (127 Abiturient*innen in Englisch-Leistungskursen und 324 Studienanfänger*innen in englischsprachigen Bachelor-Studiengängen) erreicht werden, die 674 integrierte Schreibprodukte verfassten. Damit wurde das ursprüngliche Ziel von N = 500 Personen nicht ganz erreicht, aber dennoch eine akzeptable Analysegrundlage geschaffen.

5.1 Beteiligte Promovierende:

Promovierende	Geschlecht (M/W/D)	Promotionsstatus (laufend, beendet, abgebrochen)	Promotionsbeginn und (ggf.) -ende: MM/JJJJ – MM/JJJJ	Finanzierung im Rahmen des Projektes MM/JJJJ – MM/JJJJ
Delgado Osorio, Ximena	W	beendet	01.2020–10.2024	01.2020–12.2022 02.2024–10.2024
Koval, Valeriia	W	laufend	03.2020 – 09.2026	03.2020–07.2023 03.2024-11.2025