Verarbeitung von Hinweisreizen beim Beurteilen von mündlichen Sprachproben von Schülerinnen und Schülern durch Hochschullehrende im Fach Französisch

Stéfanie Witzigmann · Steffi Sachse

Online publiziert: 16. Juni 2020
© Der/die Autor(en) 2020

Zusammenfassung Mündliche Sprachproduktionen stellen eine wichtige Zielsetzung des Fremdsprachenunterrichts dar. Demnach bilden adäquate Beurteilungen mündlicher Sprachproduktionen die entscheidenden Prämissen eines unterstützen Fremdsprachenunterrichts und gehören zu den diagnostischen Kompetenzen von Fremdsprachenlehrkräften. Mündliche Produktionen lassen sich auf inhaltlicher und sprachlicher Ebene (z. B. hinsichtlich Aussprache, Wortschatz, aber auch pragmatischer Elemente) beschreiben. Diese sprachlichen Merkmale können in unterschiedlichem Maße in die Beurteilung mündlicher Sprachproben einfließen. Um diesen Einfluss auf die Genese von diagnostischen Urteilen zu überprüfen, wurden authentische Videovignetten von einer Expertengruppe (Hochschulprofessor*innen bzw. Hochschuldozent*innen der Fachdidaktik Französisch) beurteilt. Es erfolgte ein erstes holistisches Gesamturteil, im Anschluss erhielten die Proband*innen ein analytisches Bewertungsraster und wurden danach erneut um ein holistisches Gesamturteil gebeten. Dabei wurde untersucht, zu welchem Anteil Einzelmerkmale der Sprachproduktionen von den Beurteilenden zu einem Gesamturteil integriert wurden. Zusätzlich wurde exploriert, inwiefern der berufliche Hintergrund die Urteile beeinflusst. Die Ergebnisse lassen hohe Spannweiten zwischen den Beurteilenden bei denselben Videovignetten erkennen. Die Beurteilerübereinstimmungen sind jedoch sehr hoch und belegen, dass die einzelnen sprachlichen Merkmale zuverlässig bewertet werden. Die Ergebnisse der Dominanzanalysen zeigen, dass die sprachlichen Merkmale unterschiedlich stark in die Urteilsbildung einfließen, bei der Expertengruppe aber ähnliche Varianzaufklärungen in beiden holistischen Urteilen belegt werden konnten und sich beide Urteilsmodi nur gering unterscheiden. Der berufliche Hintergrund der Expertengruppe scheint mit einer unterschiedlichen Gewichtung der Einzelmerkmale im Gesamturteil einherzugehen. Diese Ergebnisse können als

S. Witzigmann (✉) · S. Sachse
Pädagogische Hochschule Heidelberg, Keplerstraße 87, 69120 Heidelberg, Deutschland
E-Mail: witzigmann@ph-heidelberg.de

Grundlage für weitere Modellierungen in anderen Gruppen (z.B. erfahrene Lehrkräfte, Lehramtsstudierende) dienen.

Schlüsselwörter Diagnostische Kompetenz · Mündliche Sprachproben · Authentische Videovignetten · Holistische und analytische Bewertungen

Processing of cues in the assessment of oral language samples from pupils by university teachers in French

Abstract Oral language productions are an important objective of foreign language teaching. Adequate assessments of oral productions are crucial premises of supportive foreign language teaching and are part of the diagnostic skills of foreign language teachers. Oral productions can be described at the levels of content and language (e.g. pronunciation, vocabulary but also pragmatic elements). These linguistic features may be considered to varying degrees in the assessment of oral language samples. To examine this influence on the genesis of diagnostic judgements, authentic video vignettes were assessed by a group of experts (university professors and lecturers of French didactics). A first holistic assessment was carried out, followed by an analytical rating grid and another holistic reassessment. This allowed examining to what extent features of the language productions were integrated by the raters into an overall assessment. Additionally, it was explored whether the professional background of the raters exerts an influence on the judgements. The results reveal high ranges between the experts for the same video vignettes. However, the interrater reliability is very high and proves that individual linguistic features are reliably assessed. The results of the dominance analysis show that the linguistic features are incorporated to varying degrees into the formation of judgements. The high percentages of explained variance in both holistic assessments demonstrate nevertheless that the two assessments modes differ only slightly within the group of experts. The professional background of the experts suggests a different weighting of the individual characteristics in the overall assessment. These results can serve as a basis for further investigation of other groups (experienced teachers, student teachers).

Keywords Diagnostic competences · Speech samples · Authentic video vignettes · Holistic and analytic scoring

1 Einleitung

Dem Sprechen wird im schulischen Fremdsprachenunterricht ein besonderer Stellenwert zugeschrieben, da hier die Basis zur „kommunikativen Handlungsfähigkeit“ (Zydati 2005) gelegt wird. Fremdsprachenlernen soll dazu befähigen, sich in authentischen Situationen sprachlich korrekt zu äußern und situationsangemessen kommunikativ zu handeln. Der Gemeinsame Europäische Referenzrahmen für Sprachen (GER) (Europarat 2001) zeigt deutlich, dass das Sprechen eine zentrale Rolle im Fremdsprachenunterricht einzunehmen hat. Die zunehmende Aufwer-

tung von Mündlichkeit (u. a. Becker-Mrotzek 2012) in der gegenwärtigen Ära des „(neo)kommunikativen“ (Meißner 2005) Fremdsprachenunterrichts und in den Bildungsstandards (Decke-Cornill und Küster 2014; Fäcke 2017; Thiele 2012) setzt jedoch voraus, dass Lehrpersonen ihre Schüler*innen entsprechend beurteilen und fördern können (Eriksson 2012).

Mündliche Produktionen sind allerdings aufgrund ihrer Komplexität schwierig zu beurteilen (Henrici et al. 2003; Luoma 2004; Hu und Leupold 2008; O’Sullivan 2012). Beim Sprechen selbst handelt es sich um einen sehr schnell ablaufenden Prozess, die Reziprozität in einer Interaktion ist nur bedingt planbar und gesprochene Wörter sind sehr kurzlebig (Bygate 1987; Luoma 2004; Müller-Hartmann und Schöcker-von Ditzfurth 2011; Nieweler 2010). Zahlreiche interne und externe Faktoren, die das Urteil beeinflussen, stellen eine Herausforderung beim Beurteilen mündlicher Leistungen dar (Chuang 2009; Eckes 2010; Luoma 2004; Tesch und Grotjahn 2010). Wahrnehmbare Faktoren sind u. a. inhaltliche und sprachliche Merkmale (z. B. Aussprache, Umfang und Differenziertheit des Wortschatzes, grammatikalische Korrektheit, aber auch pragmatische Elemente), die als Hinweisreize aufgefasst werden können. Es wird angenommen, dass diese Hinweisreize in unterschiedlichem Maße in die Beurteilung einfließen. Ähnlich wie bei Schreibleistungen, zählen zu weiteren möglichen Einflussfaktoren Personencharakteristika der Beurteilenden (z. B. Erfahrung, beruflicher Hintergrund, Strenge/Milde und Einstellungen), Merkmale der Aufgaben (z. B. Art der Testdurchführung, Schwierigkeit der Aufgabe und Art des Inputmaterials), Merkmale der Lernenden (z. B. sprachliche Fähigkeiten, Alter, Geschlecht und kultureller Hintergrund) oder die verwendeten Beurteilungskriterien und -skalen (z. B. Uni-, Bi-, Multi-Level-Ansatz oder holistische bzw. analytische Beurteilung) (Tesch und Grotjahn 2010; Grotjahn und Kleppin 2017).

In den Studien, die diagnostische Kompetenzen von Lehrpersonen in den Blick nehmen (s. Abschn. 2.2), wird kaum auf den höchst komplexen Prozess der Beurteilung von Sprechleistungen eingegangen und entsprechend nicht untersucht, welche direkten bzw. indirekten Faktoren die Beurteilungen beeinflussen.

Die vorliegende Studie fokussiert die Genese von Urteilen bei Expert*innen des Französischen über mündliche Sprachproduktionen von Schüler*innen. Dabei wird untersucht, in welchem Ausmaß die Merkmale der Sprachproduktionen von den Beurteilenden zu einem Gesamturteil integriert werden. Der besondere Ansatz besteht darin, die implizite Merkmalsnutzung bei holistischen Urteilen und die explizite bei analytischen Urteilen in Bezug zu setzen. Zusätzlich wird exploriert, inwiefern sich die Urteile bei Beurteilenden mit unterschiedlichem beruflichem Hintergrund unterscheiden.

2 Theoretischer Hintergrund

2.1 Erforschung diagnostischer Kompetenzen

In den letzten drei Jahrzehnten wurden die diagnostischen Kompetenzen von Lehrkräften vermehrt untersucht. Dabei wurde der Fokus häufig auf die Urteilsgenauigkeit, d. h. die Übereinstimmung eines Urteils der Lehrkraft mit einer objektiven

Abb. 1 Erklärung der diagnostischen Urteile von Lehrkräften durch kognitive Modellierung. (In Anlehnung an Loibl et al. 2020)

Bewertung der Merkmale der Schüler*in gelegt und untersucht, welche Faktoren Einfluss auf diese Urteilsgenauigkeit haben (s. z. B. die Meta-Studien von Hoge und Coladarsi 1989; Südkamp et al. 2012). Die Frage nach der Genese von Urteilen ist hingegen weniger erforscht (Herppich et al. 2018; Leuders et al. 2018; Schrader 2009; Spinath 2005).

Loibl et al. (2020) entwickelten das Rahmenmodell DiaCoM (Explaining Teachers' Diagnostic Judgements by Cognitive Modeling, s. Abb. 1). Ziel von DiaCoM ist es, einen Untersuchungs- und Interpretationsrahmen für die Analyse kognitiver Prozesse beim Beurteilen von Lehrpersonen zu bieten. In der vorliegenden Studie werden die einzelnen Elemente des Rahmenmodells in folgender Weise spezifiziert und untersucht: Fremdsprachenlehrkräfte nehmen eine Sprachproduktion von Schüler*innen wahr und generieren dazu ein Gesamturteil. Wir gehen davon aus, dass sie hierbei einzelne Merkmale der Sprachproduktion (z.B. Aussprache, Verständlichkeit oder Kommunikationsstrategien) wahrnehmen und die Beurteilung der Einzelmerkmale zu einem Gesamturteil integrieren. Dies kann implizit (holistisch) oder explizit (nach expliziter Bewertung der Einzelmerkmale) geschehen. Aufgrund unterschiedlicher Voraussetzungen im Wissen oder beruflichen Vorerfahrungen sind interindividuelle Unterschiede bei der Urteilsgenese zu erwarten.

2.2 Diagnostische Kompetenz im Bereich der mündlichen Sprachproduktion

Auch wenn aktuell das Thema der Beurteilung mündlicher Sprachproduktion seitens der Lehrkräfte zunehmend aufgegriffen wird (u.a. Behr und Kierepka 2002; Diehr 2005, 2006a, b; Diehr und Frisch 2008; Drese 2008; Institut zur Qualitätsentwicklung im Bildungswesen 2019; Neumüller 2010; Nieweler 2010; Vogt 2006), finden sich nur vereinzelt Studien, die die dafür notwendige diagnostische Kompetenzen in den Blick nehmen (Diehr und Frisch 2008; Drese 2008; Hochstetter 2011; Weh 2013).

Erste Hinweise auf die Charakteristiken der beurteilenden Lehrkräfte liefern die Studien von Drese (2008) und Hochstetter (2011), die sich thematisch gezielt mit der Beobachtung mündlicher Sprachproduktion und Sprachrezeption auseinandersetzen. So nennt Drese (2008) für Lehrpersonen fachliches und fachdidaktisches Wissen, Berufserfahrung in Form von gelungenem Klassenmanagement sowie die Fähigkeit konstruktives Feedback an die Lernenden zu richten, als wichtigste Kriterien, um mündliche Sprechleistungen im Unterricht realisieren und gezielt einschätzen zu können. Bei der Studie von Hochstetter (2011) werden die Personencharakteristika der Lehrkräfte unter mehreren Aspekten analysiert. Sie geht der Frage nach, ob Lehrkräfte die Leistungen identischer Lernender kongruent einschätzen und überprüft die Validität eines eingesetzten Diagnoseinstruments (Verständnis der Beobachtungskategorien). Sechs von acht Beobachtungskategorien wurden in ihrer Studie als valide beurteilt. Hochstetter (2011) verzeichnete eine hohe Interrater-Reliabilität, die jedoch bei den leistungsschwächeren Schüler*innen geringer ausfiel.

Diese meist qualitativ angelegten Studien liefern erste Hinweise darauf, welche Situationen und Personencharakteristika für die Diagnose mündlicher Sprachproduktion bei beurteilenden Lehrpersonen vorliegen, doch über dabei ablaufende Urteilsprozesse ist bisher wenig bekannt. Es fehlen Studien, die alle vier Komponenten der diagnostischen Kompetenzen nach Loibl et al. 2020 (s. Abb. 1) in den Blick nehmen und sich auf empirische Befunde stützen.

2.3 Beurteilung mündlicher Sprachproduktionen

2.3.1 Einfluss der Situation auf den Beurteilungsprozess

Im Hinblick auf die Beurteilung von mündlichen Sprachproduktionen kann die Struktur der Beurteilungsskala einen entscheidenden Beitrag zur Genese von Urteilsprozessen leisten. Allgemein kann zwischen einer holistischen oder analytischen Beurteilung unterschieden werden (vgl. Bachman und Palmer 2010; Chuang 2009; Grotjahn und Kleppin 2008, 2017; Harsch und Martin 2013).

Bei der holistischen Beurteilung wird ein Gesamturteil abgegeben, das die Leistung der Lernenden insgesamt betrachtet und auf die explizite Bewertung einzelner Teildimensionen verzichtet. In der Regel werden die Beurteilenden unmittelbar nach der Sichtung des Sprachprodukts angehalten, die Sprachproduktion in ihrer Gesamtheit relativ schnell und auf Basis des gewonnenen Eindrucks zu beurteilen (Hinger und Stadler 2018; Knoch 2011a).

Bei der analytischen Kodierung wird hingegen das Sprachkonstrukt differenziert in seine einzelnen Teildimensionen gegliedert. Diese Einzelmerkmale (sprachliche oder inhaltliche Aspekte) werden getrennt voneinander bewertet. Diese können als Bewertungsskalen bzw. -raster realisiert werden, die je nach Leistungsstand der Lernenden, eingesetztem Aufgabenformat (Merkmale müssen die gestellte Aufgabe abbilden können) oder zu messendem Sprachkonstrukt (mündliche oder schriftliche Produktion) unterschiedliche Formen annehmen können (vgl. Brookhart 2013; Grotjahn und Kleppin 2017).

Um das Bilden eines Gesamturteils nach einer analytischen Bewertung vom ersten holistischen Gesamturteil zu unterscheiden, verwenden Grotjahn und Kleppin

(2017) den Begriff der globalen Bewertung¹ und Porsch (2010) den Begriff Globalurteil. Im Sinne eines aggregierten Urteils, beruhend auf einer zuvor erfolgten analytischen Bewertung, werden wir den Begriff Globalurteil ebenfalls in unserer Studie verwenden.

Empirische Studien, vor allem im Bereich der Schreibkompetenzen, bestätigen, dass das Sprachkonstrukt bei beiden Urteilsmodi (holistisch und analytisch) reliabel beurteilt werden kann (vgl. u. a. Porsch 2010; Schipolowski und Böhme 2016). Werden holistische Urteile in Form eines Gesamteindrucks mit analytischen Urteilen in Bezug gesetzt, sind sie für das Sprachkonstrukt nicht nur reliabel (vgl. Chuang 2009; Goulden 1994; Metruk 2018; Ounis 2017), sondern haben den Vorteil, die implizite Nutzung der einzelnen sprachlichen Merkmale aufzeigen zu können.

Die akkurate Beurteilung von mündlichen Sprachproduktionen sollte möglichst auf der Basis von konstruktrelevanten Merkmalen erfolgen. Im Bereich der mündlichen Produktion lassen sich diese Merkmale im Bereich der linguistischen Kompetenz, der strategischen Kompetenz, des Textwissens, des pragmatischen Wissens und des soziologischen Wissens identifizieren (Europarat 2001; Hutterli 2011). Nachfolgend werden sechs Merkmale auf der linguistischen und strategischen Kompetenzebene (Phonologie, Lexik, Grammatik, Flüssigkeit, Verständlichkeit und Kommunikationsstrategien), die für die Beurteilung herangezogen werden können, beschrieben.

Die *Phonologie* umfasst u. a. die Beherrschung des Lautsystems einer Sprache.

Die *Lexik* umfasst die Kenntnis des Vokabulars einer Sprache. Polyseme Wörter können auf dieser Ebene Schwierigkeiten in der Beurteilung von Einzelwörtern darstellen. In der Diagnostik der Sprachkompetenz sollte sowohl der Wortschatzumfang als auch das Wortschatzspektrum (bezogen auf die Kommunikationsaufgabe) bedacht werden.

Grammatik bezieht sich auf die Kenntnis der Regelemente, die sich sowohl auf einzelne Wörter (Morphologie) wie auch auf Sätze (Syntax) applizieren lässt. Bei der Beurteilung sollte die grammatikalische Korrektheit je nach Kommunikationsabsicht bzw. Diskursform (Diehr 2005) individuell bewertet werden (geplante Sprechproduktion versus spontane Äußerungen, welche Ellipsen oder Anakoluthen durchaus dulden).

Für die *Flüssigkeit* des Sprechens können sowohl gefüllte (mit Geräuschen wie „emh“) oder ungefüllte (stille) Pausen oder auch nur die Sprechgeschwindigkeit (*speed fluency*) berücksichtigt werden (Fulcher 2015; Tavakoli und Skehan 2005). Bei der Beurteilung von Flüssigkeit (Redefluss) soll in unserer Studie ausschließlich von der Sprechgeschwindigkeit ausgegangen werden.

Die *Verständlichkeit* versteht sich als eine gelungene Verbindung zwischen mehreren linguistischen und strategischen Kompetenzen, die den Inhalt einer Sprechabsicht vollständig und korrekt zu verstehen gibt. Die Beurteilung der Verständlichkeit soll in unserer Studie sowohl linguistische wie auch paralinguistische Elemente inkludieren, da diese im Sinne einer Botschaftsvermittlung verstanden werden sollen.

¹ Nicht zu verwechseln mit dem Terminus „Globaleindruck“, der von Porsch (2010) als holistische Auswertung vor dem Einsetzen der analytischen Kriterien verwendet wird.

Kommunikationsstrategien werden im Kontext der Fremdsprache auf mehreren Ebenen vom Lernenden aktiviert, um die Anforderungen einer Aufgabe zu bewältigen. Paralinguistische Strategien wie Gestik/Mimik, Lautmalerei oder prosodische Mittel (Lautstärke, Stimmhöhe) können entweder als Kompensationsmittel eingesetzt werden oder um eine Aussage zu verstärken bzw. visuell zu untermauern als kommunikationsfördernde Strategien.

2.3.2 Einfluss der Personencharakteristika auf den Beurteilungsprozess

Personencharakteristika umfassen die Eigenschaften von beurteilenden Personen wie Alter, Geschlecht, Einstellungen, beruflicher Hintergrund, Wissen oder Erfahrung und können je nach Diagnosesituation einen Einfluss auf die Urteile haben (vgl. u. a. Eckes 2010, 2015; Shaw 2007). Man spricht dabei von einer „beurteilergestützten Leistungsmessung“ (*rater-mediated assessment*) (vgl. u. a. Eckes 2010, 2015; Knoch 2011a). Die besondere Herausforderung für die Beurteilenden besteht darin, ausschließlich die Ausprägungen der Sprachkompetenzen der Lernenden akkurat wahrzunehmen, interpretieren und beurteilen zu können, möglichst ohne sich von einer Reihe von Urteilstendenzen (z. B. Tendenz zur Strenge/Milde, zentrale Tendenz oder Halo-Effekte), Schwierigkeiten der Aufgaben/Kriterien, Strukturen der Ratingskala, persönlichen Einstellungen oder Gemütszuständen bzw. motivationalen Faktoren (Caban 2003; Eckes 2010; Lenske 2016) beeinflussen zu lassen. Ein Halo-Effekt ist dann gegeben, wenn ein konstruktirrelevantes Merkmal (sei es positiv oder negativ), wie bspw. die ausgestrahlte Sympathie eines Lernenden, die Beurteilung anderer relevanter Aspekte der mündlichen Sprachproduktion beeinflusst. Faktoren wie Strenge/Milde (entweder generell oder nur bei bestimmten Personen) sind schwierig zu verändern, da es sich um relativ stabile Persönlichkeitsmerkmale handelt (Eckes 2015; Knoch 2011b).

Im Bereich der Fremdsprachenforschung insbesondere bei mündlichen Beurteilungen konnten frühere Studien den Einfluss der Rater auf die Testergebnisse (u. a. Bachman et al. 1995; Brown 1995; Caban 2003; Lumley und McNamara 1995) aufzeigen. Einen entscheidenden Einfluss übte dabei der (berufliche) Hintergrund bzw. das Fachwissen der Bewertenden aus. Kolb (2011) konnte bspw. zeigen, dass Sekundarschullehrkräften im Fach Englisch andere methodische Arbeitsformen einsetzen als Primarschullehrkräfte. So liegt der Schwerpunkt in der Primarstufe überwiegend auf mündlichen Arbeitsformen (und mündlichen Lernstandskontrollen) während sie in der Sekundarstufe einen geringeren Stellenwert einnehmen (nur 9 % der befragten 270 Lehrkräfte setzten am Gymnasium mündliche Lernstandskontrollen ein). Zudem stellte Kolb (2011) fest, dass Primarstufenlehrkräfte weniger auf formale Aspekte (wie Grammatik oder Vokabelarbeit) der Sprache achten als Sekundarschullehrkräfte (für ähnliche Ergebnisse s. Burwitz-Melzer und Legutke 2004; Drese 2004).

Inwiefern sich die Beurteilenden in ihren Urteilen unterscheiden, kann durch die Beurteilerreliabilität erfasst werden. Dabei wird zwischen Inter- und Intrarater-Reliabilität unterschieden. Die Interrater-Reliabilität bezeichnet die Übereinstimmung von Urteilen zwischen unterschiedlichen Beurteilenden. Eine hinreichende Interrater-Reliabilität ist dann gegeben, wenn die Unterschiede zwischen den Urteilen möglichst gering ausfallen oder die Lernenden zumindest in eine weitgehend ana-

loge Rangfolge gebracht werden (*relative consistency*). Von Intrarater-Reliabilität wird hingegen gesprochen, wenn ein und derselbe Beurteilende über alle mündliche Produktionen hinweg in konsistenter Weise die Merkmale interpretiert und in vergleichbarer Weise streng/mild urteilt (Gwet 2014; LeBreton und Senter 2008; Wirtz und Caspar 2002).

3 Fragestellungen und Hypothesen

Die vorliegende Studie untersucht die Genese von Urteilen von Lehrkräften über mündliche Sprachproduktionen von Schüler*innen. Dabei wird untersucht, in welchem Ausmaß Merkmale der Sprachproduktionen für die Bildung eines Gesamturteils genutzt werden. Um hier idealtypische Urteilsgenesen abzubilden, wird die Studie mit fachdidaktischen Expert*innen durchgeführt. Es stellt sich die Frage, inwiefern die Expert*innen die Sprachproduktionen in ihrer Gesamtheit, aber auch die einzelnen sprachlichen Merkmale der Sprachproduktion übereinstimmend bewerten. Daher lautet die erste Forschungsfrage (FF1): Wie stark variieren die Expert*innenurteile und wie hoch ist die Übereinstimmung zwischen den Expert*innen?

Im Fokus der vorliegenden Studie steht der Zusammenhang zwischen den Urteilen über Einzelmerkmale und dem Gesamturteil. Der Ansatz der vorliegenden Studie besteht dabei darin, die implizite Nutzung der Einzelmerkmalsurteile bei holistischen Gesamturteilen und die explizite Nutzung bei analytischen Gesamturteilen in Bezug zu setzen. Daher lautet die zweite Forschungsfrage (FF2): (a) Wie stark fließen die Urteile über sprachliche Einzelmerkmale in das Gesamturteil ein? (b) Ist dieser Einfluss unterschiedlich, je nachdem ob das Gesamturteil vor oder nach den Einzelurteilen erfolgt?

Zusätzlich wird exploriert, inwiefern sich die Urteile der Expert*innen mit einem beruflichen Schwerpunkt in der Primarstufe und/oder der Sekundarstufe unterscheiden. Diese Unterscheidung der Expert*innen soll aufgrund der geringen Fallzahlen jedoch lediglich auf deskriptiver Ebene betrachtet werden.

4 Methode

4.1 Stichprobe

Bei den Proband*innen der Studie handelte es sich um Fachdidaktiker*innen ($N=13$) der Zielsprache Französisch, die sich durch ihr bereichsspezifisches Wissen innerhalb der Domäne „Didaktik des Französischen“ (Professor*innen und Dozent*innen an Hochschulen oder Seminarleiter*innen der zweiten Ausbildungsphase) kennzeichnen (s. Bromme 2001, 2014; Gruber 2010; Krauss und Bruckmaier 2014). Davon waren sechs männlich (46,15 %) und sieben weiblich (53,85 %). Fast die Hälfte der Proband*innen waren zwischen 40 und 49 Jahre alt (46 %), fünf davon zwischen 50–59 Jahre alt (38 %) und 15 % zwischen 60 und 69 Jahren. Alle Fachdidaktiker*innen der vorliegenden Studie zeichnen sich durch hervorragende Fran-

zösischenkenntnisse aus, sind in der fachdidaktischen Lehre mit wissenschaftlichem Bezug verortet, haben eine abgeschlossene Lehramtsausbildung (mit 1. und 2. Staatsexamen) und weisen eine mehrjährige Schulpraxis (>3 Jahre) auf. Diese Gruppe wird demzufolge als Expert*innen bezeichnet. Zwar verfügen Fachdidaktiker*innen meist über überdurchschnittliche sprachliche und z. T. fachdidaktische Kenntnisse, doch die Hochschullehre bzw. das Studium kann sich ggfs. stark strukturell und inhaltlich bei der Lehramtsausbildung unterscheiden. In Baden-Württemberg findet bspw. die Lehramtsausbildung für die Primarstufe und Sekundarstufe I an Pädagogischen Hochschulen statt, während die Lehramtsausbildung für die Sekundarstufe II an Universitäten erfolgt. So erhalten Lehramtsstudierende an Pädagogischen Hochschulen in Baden-Württemberg für die Lehramtsausbildung in der Primarstufe im Durchschnitt ca. doppelt so viele fachdidaktische Lehrveranstaltungen als Lehramtsstudierende an der Universität². Doch auch bei den Lernzielen und Vermittlungspraxen werden von beiden Hochschulgruppen unterschiedliche Schwerpunkte gesetzt. Die Hochschullehre adressiert die Bedürfnisse der zukünftigen Zielgruppe der Schüler*innen in den jeweiligen Bildungsgängen und passt ihre Lehre entsprechend an. Somit sind die fachdidaktischen Ausbildungsschwerpunkte, wie auch Berufserfahrungen der einzelnen Fachdidaktiker*innen in der vorliegenden Studie strukturell unterschiedlich, so dass eine Unterteilung zwischen den Lehrämtern (Primar- versus Sekundarstufe) durchaus gegeben ist. Der berufliche Hintergrund der Expert*innen ermöglichte die Identifizierung von sechs Expert*innen im Primarbereich sowie sieben im Sekundarbereich.

4.2 Untersuchungsdesign und Operationalisierung

Da mündliche Sprachproduktionen nur der direkten Beobachtung unterliegen, wurde die Diagnosesituation anhand von videografierten Sprachproben (Stimulusmaterial) realisiert. Videovignetten erlauben, im Gegensatz zu Audioaufnahmen, die Wahrnehmung von paralinguistischen Mitteln wie Gestik und Mimik. Der Informationsgehalt der audiovisuellen Sprachproben ist somit durch die Anreicherung mit inhaltsbezogenen Merkmalen wie Gestik und Mimik intensiviert. Vier Schulklassen an verschiedenen Schulen (2 Grundschulklassen – zum Ende der 4. Klasse und 2 Gymnasialklassen zu Beginn der 5. Klasse) wurden ausgewählt. Die Lernenden wurden einzeln videografiert und alle mit der gleichen Aufgabenstellung konfrontiert: In einem ersten Teil sollten sie – aufgefordert durch vier Fragen einer Muttersprachlerin – sich selbst, ihre Familie, ihre Hobbies und ihr Lieblingstier vorstellen. Im zweiten Teil sollten sie einen Comic beschreiben (acht Bilder mit einer kurzen Begebenheit zwischen dem Hauptprotagonisten (einem Hund) und anderen Hunden). Insgesamt wurden neun Videovignetten selektiert (fünf Jungen und vier Mädchen im Alter zwischen 10 und 11 Jahren), die jeweils ca. 2–3 min lang sind und zwei unterschiedliche sprachliche Anforderungen enthalten. Bei den Sprachproben

² Werden bspw. die Lehramtsausbildung an der Universität und der Pädagogischen Hochschule in Heidelberg verglichen, so beinhaltet das Studium (Bachelor und Master) im gewählten Unterrichtsfach für die Primarstufe an der PH insgesamt 14 ECTS an fachdidaktischen Lehrveranstaltungen während die Fachdidaktik im selben Fach für die Sekundarstufe II insgesamt 6 ECTS umfasst.

	sehr gute					sehr schlechte				
Aussprache	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
	kaum differenziert					sehr vorhanden				
Wortschatz	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩

Abb. 2 Sprachliche Merkmale, Beispielitems

handelt es sich um echte Sprechsituationen. Die Lernenden wurden nicht spezifisch auf die Fragestellungen vorbereitet und kannten den Comic nicht.

Das Stimulusmaterial wurde den Expert*innen vorgelegt. Die Informationen über die Gesamtgruppe der Schüler*innen waren für alle Proband*innen dieselben, bezogen auf die Authentizität der Sprachproben (keine Vorbereitung), dem Alter der Proband*innen (10–11 Jahre) und den Lernjahren in der Fremdsprache (2 Jahre, entweder Grundschule oder Gymnasium). Die Aufgabe der Expert*innen bestand darin, in einem ersten Durchgang jeweils eine Videovignette zu sichten und unmittelbar im Anschluss ein holistisches Gesamturteil (GU1) auf einer Skala von 0–100 % (100 % = perfekte Leistung) abzugeben (T1). Diese ersten Gesamturteile wurden für die nächsten acht Sprachproduktionen wiederholt.

Direkt im Anschluss wurden in einem zweiten Durchgang dieselben Videovignetten erneut präsentiert, jedoch erhielten die Expert*innen hierfür ein Beurteilungsraster (für die Merkmale s. Abschn. 2.3.2). Für jedes Merkmal lag eine numerische 10-stufige Skala vor, auf der nur der Anfang- und Endpunkt verbal markiert wurde (s. Abb. 2). Nach jeder Videovignette und der analytischen Beurteilung wurden die Expert*innen erneut aufgefordert, eine Gesamtbeurteilung (GU2), hier als sogenannte globale Beurteilung, auf der 0–100 % Skala abzugeben (T2).

4.3 Auswertungsstrategien

Die Daten wurden mittels SPSS für Windows (Version 25.0) und des open-source Statistikprogramms R (R Core-Team 2018) mit der Desktop-Plattform R-Studio (RStudio, Inc. 2018) analysiert.

Von jedem der 13 Expert*innen wurden neun Videovignetten beurteilt. Die Datenstruktur enthält Beurteilende, innerhalb derer die Stimuli/Sprachproduktionen geclustert sind. Es handelt sich somit nicht um 117 (13*9) unabhängige Datenpunkte, sondern um ein repeated-measurement Design mit $n=13$ und einen within-Faktor mit 9 Stufen. Entsprechend wurden alle Analysen für jeden Beurteilenden getrennt über alle Stimuli hinweg durchgeführt und anschließend gemittelt.

Mit Blick auf die erste Fragestellung wurden deskriptive Statistiken berechnet. Zur Klärung der großen Varianz zwischen den Gesamturteilen der einzelnen Expert*innen wurden Intra-Klassen-Korrelationen (Wirtz und Caspar 2002) berechnet. In dieser Studie wurde der unjustierte (*absolute*), zweifaktorielle Intraklassenkoeffizient ($ICC_{2,1}$) verwendet, da alle neun Videovignetten von allen Expert*innen beurteilt wurden und die Auswahl der Expert*innen zufällig erfolgte.

Zur Beantwortung der zweiten Fragestellung, in welchem Maß die Einzelmerkmale in das Gesamturteil einfließen, wurden zunächst multiple lineare Regressionen durchgeführt, um die Beziehungen zwischen den Hinweisreizen und den abhängigen Urteilen zu untersuchen. Zwar konnten signifikante Zusammenhänge zu einigen Hinweisreizen festgestellt werden, doch die Aussagekraft der Beta-Werte bei der multiplen linearen Regression blieb eingeschränkt. Um dem nachzugehen, wurden zunächst Korrelationen zwischen den Hinweisreizen und dem zweiten Gesamturteil berechnet, die teilweise sehr hoch ausfielen (s. Tab. 3). Laut Field (2018) sind Korrelationswerte über 0,8 ein Anzeichen für Multikollinearität. Durch den Einsatz der Korrelationsmatrix (Korrelationen der unabhängigen Variable) und vor allem der Toleranzwerte (*TOL*) (s. Tab. 4) ließ sich Multikollinearität feststellen. Ein Wert von $TOL < 0,2$ deutet auf eine starke Multikollinearität hin. Um dem Rechnung zu tragen, wurden Dominanzanalysen (Budescu 1993) durchgeführt. Die Dominanzanalyse hat den Vorteil, dass nicht zwingend auf die Interkorrelationen der Prädiktoren geachtet werden muss, weil alle Prädiktoren im gleichen Maß berücksichtigt werden. Bei diesem Verfahren werden die Prädiktoren in sämtlichen möglichen Konstellationen miteinander verglichen, um herauszufinden, welcher Prädiktor dominant ist. Dabei werden die Resultate der multiplen Regression genutzt, um die Dominanz einer Variablen über die andere zu bestimmen. Für jede vorangehende Variablenkombination wird für eine fokale Variable die zusätzliche Varianzaufklärung berichtet (ΔR^2). Die Varianzaufklärungen für eine fokale Variable werden über die verschiedenen Konstellationen mit allen möglichen vorangehenden Variablenkombinationen gemittelt. Gegenüber standardisierten Beta-Werten (vgl. Grömping 2015) sollten generelle Dominanzwerte präferiert werden, da sie als valide Indikatoren von individueller Variablenbedeutung, gerade bei Interkorrelationen zwischen den unabhängigen Variablen, gelten (Tonidandel und LeBreton 2011; Thomas et al. 2014). Die Effekte können ähnlich wie Cohens *R*-Quadrat interpretiert werden: $\geq 0,02$ kleiner Effekt, $\geq 0,13$ moderater Effekt und $\geq 0,26$ starker Effekt. Als Prädiktoren gingen die beurteilten sprachlichen Einzelmerkmale ein. Kriterien waren das zweite Gesamturteil und das erste Gesamturteil. Dadurch kann nicht nur die Gewichtung der Prädiktoren beim Globalurteil (GU2) nach expliziter Bewertung der Einzelmerkmale untersucht werden, sondern ebenfalls die Bedeutsamkeit der vorgegebenen Einzelmerkmale für die implizite (holistische) Erstbeurteilung (GU1). Datengrundlage jeder Dominanzanalyse bildeten die neun Videovignetten.

5 Ergebnisse

5.1 FF1: Wie stark variieren die Expert*innenurteile und wie hoch ist die Übereinstimmung zwischen den Expert*innen beim Gesamturteil und der Beurteilung der Sprachmerkmale?

Die deskriptiven Statistiken der erhobenen Gesamturteile sind in Tab. 1 dargestellt und enthalten die Mittelwerte und Standardabweichungen für beide Urteile (GU1 und GU2). Die hohe Standardabweichung von 18,66 sowie der Range (min=25,

Tab. 1 Mittelwerte, Standardabweichungen, Spannweite und Rang beider Gesamturteile über $n = 13$ Expert*innen

	Geschlecht	M	SD	Min	Max	Rang
<i>GU1</i>						
SID-1	m	52,08	17,47	20	80	6
SID-2	w	70,69	18,19	25	98	3
SID-3	m	20,00	10,02	5	35	9
SID-4	w	63,07	18,65	40	95	4
SID-5	m	29,92	15,38	5	60	8
SID-6	m	48,69	17,02	20	72	7
SID-7	w	75,69	13,17	50	100	2
SID-8	w	62,23	18,66	25	90	5
SID-9	m	82,00	13,71	40	95	1
<i>GU2</i>						
SID-1	m	51,69	18,04	20	80	6
SID-2	w	57,15	15,50	35	80	4
SID-3	m	24,53	10,45	5	40	9
SID-4	w	58,31	16,82	30	85	3
SID-5	m	32,77	14,05	15	58	8
SID-6	m	44,77	16,32	20	75	7
SID-7	w	74,62	10,21	62	95	2
SID-8	w	53,31	13,60	35	77	5
SID-9	m	77,69	15,24	35	92	1

SID Sprachproduktion-ID, *GU1* erstes Gesamturteil, *GU2* zweites Gesamturteil

max = 90) für dieselbe Sprachproduktion (z. B. SID-8) zeigen, dass die Urteile teilweise hohe Spannweiten zwischen den einzelnen Expert*innen aufweisen.

Tab. 1 zeigt, dass die Lernenden beim zweiten Gesamturteil (GU2), hier nach dem analytischen Beurteilungsraster, etwas moderater eingeschätzt werden, also mit einem leichten Trend zur Mitte. Die Standardabweichungen fallen beim GU2 im Mittel geringfügig kleiner aus (SD = 14,47) als beim GU1 (SD = 15,81). In Anbetracht der angesetzten Beurteilungsspanne zwischen 0 und 100 sind die Urteile zwischen dem ersten holistischen Gesamturteil und dem zweiten Gesamturteil bei den Expert*innen vergleichbar.

Es besteht eine hohe Konsistenz zwischen dem Ersturteil (GU1) und dem Zweiturteil (GU2), da die Mittelwerte relativ konstant bleiben, außer für zwei Lernende (SID-2 und SID-8), wo offensichtlich eine Revision durch die Expert*innen stattfand.

Um diese Übereinstimmung der Urteile bei den Expert*innen bei GU1 als auch beim GU2 spezifischer zu prüfen, sind Intraklassenkoeffizienten berechnet worden (s. Tab. 2).

Die Werte des ICC liegen meist über 0,90, was bedeutet, dass die Übereinstimmung der Expert*innen über alle Hinweisreize hinweg sehr hoch ist. Lediglich bei der Aussprache fällt die Übereinstimmung etwas geringer aus (ICC = 0,83). Insgesamt lassen sich die Lernenden sowohl beim ersten holistischen Gesamturteil wie

Tab. 2 Ergebnisse der Berechnungen der Interklassenkoeffizienten ($n = 13$)

Hinweisreize	ICC (2,1)
Aussprache	ICC = 0,83 ($F = 7,69$; $p < 0,001$)
Wortschatz	ICC = 0,94 ($F = 18,60$; $p < 0,001$)
Grammatik	ICC = 0,93 ($F = 16,31$; $p < 0,001$)
Redefluss	ICC = 0,92 ($F = 18,50$; $p < 0,001$)
Verständlichkeit	ICC = 0,94 ($F = 20,67$; $p < 0,001$)
Kommunikationsstrategien	ICC = 0,97 ($F = 38,30$; $p < 0,001$)
Sprachwechsel	ICC = 0,99 ($F = 199,86$; $p < 0,001$)
GU1	ICC = 0,954 ($F = 32,95$; $p < 0,001$)
GU2	ICC = 0,946 ($F = 23,25$; $p < 0,001$)

auch bei der globalen Beurteilung weitestgehend (bis auf Rang 3 und 4, s. Tab. 1) in die gleiche Reihenfolge bringen.

5.2 FF2: (a) Wie stark fließen die Urteile über Einzelmerkmale in das Gesamturteil ein? (b) Ist dieser Einfluss unterschiedlich, je nachdem ob das Gesamturteil vor oder nach den Einzelurteilen erfolgt?

(a) Ein signifikanter Einfluss der sprachlichen Merkmale konnte zwar durch die multiple lineare Regression gezeigt werden (für das GU2: Wortschatz: $\beta = 0,465$; $p < 0,001$; Grammatikalische Korrektheit: $\beta = 0,155$; $p = 0,020$; Verständlichkeit der Produktion: $\beta = 0,279$; $p < 0,001$), doch diese Ergebnisse bleiben zur Beantwortung der Fragestellung unbefriedigend. Die korrelativen Analysen (s. Tab. 3 und 4) und vor allem die Toleranzwerte (*TOL*) lassen Multikollinearität feststellen.

Tab. 3 Korrelationen zwischen GU2 und den Hinweisreizen gemittelt über $n = 13$ Expert*innen

	Aussprache	Wortschatz	Grammatik	Redefluss	Verständlichkeit	Kommunikationsstrategien	Sprachwechsel
GU2	0,679	0,919	0,830	0,687	0,844	-0,506	0,429

GU2 zweites Gesamturteil

Tab. 4 Interkorrelationsmatrix der unabhängigen Variablen gemittelt über $n = 13$ Expert*innen

	AUS	WORT	GRAM	REDE	VERS	KOM_ST	SPR_WE	TOL
AUS	–	0,533	0,394	0,471	0,471	-0,417	0,151	0,143
WORT	0,533	–	0,627	0,417	0,582	-0,436	0,390	0,061
GRAM	0,394	0,627	–	0,280	0,501	-0,404	0,326	0,087
REDE	0,471	0,417	0,280	–	0,246	-0,366	0,175	0,092
VERS	0,471	0,582	0,501	0,246	–	-0,268	0,391	0,044
KOM_ST	-0,417	-0,436	-0,404	-0,366	-0,268	–	0,073	0,221
SPR_WE	0,151	0,390	0,326	0,175	0,391	0,073	–	0,189

AUS Aussprache, WORT Wortschatz, GRAM Grammatikalische Korrektheit, REDE Redefluss, VERS Verständlichkeit, KOM_ST Kommunikationsstrategien, SPR_WE Sprachwechsel, TOL Toleranzwerte ($1 - R^2$)

Tab. 5 Dominanzanalyse der Hinweisreize für beide Gesamturteile und Gesamtvarianzaufklärung der abhängigen Variablen gemittelt über $n = 13$ Expert*innen

	GU1	GU2
Wortschatz	0,220	0,224
Verständlichkeit	0,174	0,191
Grammatik	0,162	0,191
Kom.-Strategien	0,136	0,090
Aussprache	0,120	0,119
Redefluss	0,119	0,126
Sprachwechsel	0,070	0,059
<i>Gesamtvarianzaufklärung R^2</i>	0,964	0,984

Mittelwerte aus $n = 13$ Dominanzanalysen. In die Dominanzanalysen gehen die 7 Sprachmerkmale als Prädiktoren und das Gesamturteil (i.e. GU1, GU2) als Kriterium ein. Datengrundlage jeder Dominanzanalyse bilden die 9 Sprachproben

Aufgrund der vorhandenen Multikollinearität und zur Bestimmung der Gewichtung der Einzelmerkmale wurden deshalb Dominanzanalysen durchgeführt (s. Tab. 5).

Tab. 5 zeigt, dass beim ersten Gesamturteil der Wortschatz (mit 22%), gefolgt von Verständlichkeit (17,4%) und Grammatik (16,2%) die stärksten Prädiktoren der Varianzaufklärung für das holistische Gesamturteil bilden. Die Kommunikationsstrategien werden mit 13,6% beim ersten Gesamturteil relativ stark gewichtet, während sie beim zweiten Gesamturteil (nach expliziter Einzelmerkmalbeurteilung) kaum mehr beachtet werden. Dafür ist die Gewichtung der Verständlichkeit und grammatikalischen Korrektheit etwas höher als beim Ersturteil.

(b) Der Einfluss der Urteilsmodi auf das Gesamturteil lässt sich ebenfalls in Tab. 5 ablesen. Bis auf geringe Abweichungen (mit Ausnahme der Kommunikationsstrategien), sind die Werte der Dominanzanalyse für die einzelnen Merkmale ähnlich, unabhängig davon, ob das Gesamturteil vor oder nach den Einzelurteilen erfolgte. Auch zeigt die Gesamtvarianzaufklärung in beiden Urteilsmodi gleich hohe Werte.

Tab. 6 Dominanzanalyse der Hinweisreize für das zweite Gesamturteil (GU2) gemittelt über die Expert*innen getrennt nach Primarstufe und Sekundarstufe (Varianzaufklärungen)

GU2	Experte für Primarstufe	Experte für Sekundarstufe
Wortschatz	0,244	0,226
Verständlichkeit	0,215	0,173
Grammatik	0,175	0,177
Kommunikationsstrategien	0,095	0,067
Aussprache	0,097	0,125
Redefluss	0,114	0,166
Sprachwechsel	0,062	0,067
<i>Gesamtvarianzaufklärung R^2</i>	0,974	0,993

Mittelwerte aus $n = 6$ (Primarstufe) bzw. $n = 7$ (Sekundarstufe) Dominanzanalysen. In die Dominanzanalysen gehen die 7 Sprachmerkmale als Prädiktoren und das Gesamturteil (i.e. GU1, GU2) als Kriterium ein. Datengrundlage jeder Dominanzanalyse bilden die 9 Sprachproben

Tab. 6 stellt die Ergebnisse der Dominanzanalysen bei den Expert*innen getrennt nach beruflichem Hintergrund dar. Hierbei handelt es sich um die gemittelte Dominanz für die jeweilige Gruppe der Expert*innen.

Beim zweiten Gesamturteil nach dem Beurteilen der Einzelmerkmale bilden für die Expert*innen der Primarstufe der Wortschatz (24,4 %) und die Verständlichkeit (21,5 %) die zwei stärksten Hinweisreize der Varianzaufklärung für das Gesamturteil. Bei den Expert*innen der Sekundarstufe fließt die Verständlichkeit (17,3 %) geringer in das Gesamturteil ein. Der deskriptive Vergleich in unserer Stichprobe zeigt, dass die Expert*innen für die Sekundarstufe die Aussprache und den Redefluss höher bewerten als die Expert*innen der Primarstufe. Eine inferenzstatistische Absicherung dieses Vergleichs ist aufgrund der geringen Fallzahlen an dieser Stelle nicht angemessen.

6 Diskussion

Der situative Rahmen der vorliegenden Studie stellte authentische Sprachproben zur Beurteilung bereit. Es erfolgte die Manipulation der Urteilssituation (holistisches Gesamturteil und zweites Gesamturteil nach analytischen Einzelurteilen), um zu untersuchen, in welchem Ausmaß wesentliche linguistische Merkmale der Sprachproduktion (Grammatik, Aussprache, etc.) für die Bildung eines Gesamturteils genutzt wurden.

Im Rahmen dieser Studie konnte gezeigt werden, dass trotz unterschiedlicher Bewertungsmaßstäbe zwischen den einzelnen Beurteilenden die mündlichen Sprachproben zuverlässig bewertet wurden. Die beiden Gesamturteile (holistisches Gesamturteil und Globalurteil nach den analytischen Merkmalen) unterscheiden sich bei der Gruppe der Expert*innen nicht stark voneinander. Die einzelnen linguistischen Merkmale der Sprachproduktionen (Aussprache, Wortschatz, etc.) gehen in unterschiedlichem Ausmaß in die Bildung der Urteile ein. Die Exploration des beruflichen Hintergrundes lässt einige Unterschiede in der Gewichtung der Einzelmerkmale je nach Ausbildungstyp vermuten.

6.1 Intra- und Interrater-Reliabilität bei der Beurteilung mündlicher Sprachproben

Es liegen große Spannweiten bei den Urteilen der einzelnen Expert*innen vor. So sind sich die Expert*innen in ihren absoluten Urteilen nicht einig. Dies könnte damit zusammenhängen, dass bei den Expert*innen unterschiedliche Bezugsnormen gelten. Wurden die Leistungen der Schüler*innen mit den Leistungen der Gruppe verglichen und vor diesem Hintergrund bewertet (soziale Bezugsnorm) oder spielen bei den Expert*innen eher allgemeinere Bewertungsmaßstäbe der Zielsprache eine zentrale Rolle (sachliche bzw. kriteriale Bezugsnorm)? Auch kann vermutet werden, dass das verwendete Bewertungsinstrument (stufenlose Skala von 0 bis 100) einige Expert*innen dazu verleitete, sehr gute Leistungen entsprechend mit 100 % zu bewerten, während schlechte auch sehr niedrig beziffert wurden. Hätte man für die Beurteilungen klassische Schulnoten verwendet, wären die schlechtesten Leistungen

wohl kaum mit der Note unzureichend (6,0) bewertet worden (Birkel 2005; Hochstetter 2011). Die vorliegenden Streubreiten der Beurteilungen sind vergleichbar zu anderen Studien (vgl. u. a. Birkel 2005; Caban 2003; Chuang 2009; Porsch 2010).

Allerdings sind die Expert*innen durchaus in der Lage, die Lernenden in eine ähnliche Rangordnung zu bringen. Insgesamt werden die schwachen von den mittelstarken bis hin zu den starken Lernenden übereinstimmend differenziert. Auch sprechen die hohen Werte der Intraklassenkoeffizienten (s. Tab. 4) für eine sehr gute Interrater-Reliabilität der Expert*innen. Nur die Übereinstimmung bezüglich der Aussprache fällt etwas geringer aus ($ICC=0,83$), was laut Cicchetti (1994) noch eine sehr gute Interrater-Übereinstimmung darstellt, bei Koo und Li (2016) nur noch als gute Übereinstimmung interpretiert wird (sehr gute $\geq 0,9$). Somit zeigen die Expert*innen zwar eine große Spannweite in den einzelnen Gesamurteilen, sind sich jedoch hinsichtlich der Rangfolgen der zu beurteilenden Lernenden, auch im Hinblick auf die Bewertung der einzelnen sprachlichen Merkmale, einig und entsprechend in ihren Urteilen insgesamt konsistent.

6.2 Vergleich zwischen beiden Beurteilungsmodi

Insgesamt werden die neun mündlichen Sprachproben in beiden Urteilsmodi (holistisches Gesamturteil und Globalurteil nach dem analytischen Beurteilungsraster) von den Expert*innen sehr ähnlich beurteilt (s. Tab. 3). Erstaunlicherweise werden die Lernenden beim zweiten Globalurteil (GU2) im Mittel etwas schlechter bewertet, außer die zwei schwächsten Lernenden, die nach dem analytischen Bewertungsraster geringfügig besser bewertet werden als beim ersten Gesamturteil (vgl. ähnliche Erkenntnisse bei Birkel 2005).

Auch liegen die Varianzaufklärungen der abhängigen Variablen bei den Expert*innen in beiden Urteilsmodi in ähnlicher Höhe (s. Tab. 5). Die hohe Gesamtvarianzaufklärung für das erste holistische Gesamturteil (GU1) spricht dafür, dass die Expert*innen die sprachlichen Hinweisreize zuverlässig in den authentischen Videovignetten wahrnehmen. Dies deutet darauf hin, dass die Expert*innen die Beurteilungskriterien implizit anwenden und damit beim ersten holistischen Gesamturteil zu ähnlichen Urteilen kommen wie mit der externen Fokussierung auf bestimmte sprachliche Merkmale durch das Beurteilungsraster. Dies könnte mit dem hohen Fachwissen und fachdidaktischen Wissen der Expert*innen erklärt werden.

6.3 Personencharakteristika als Einflussvariable

Um zu untersuchen, ob es bereits in der vorliegenden Studie (trotz geringer Fallzahl) erste Hinweise auf Einflüsse des beruflichen Hintergrundes auf die Gesamturteile gibt, wurden die Expert*innen in zwei Gruppen unterteilt (s. Abschn. 4.1). Die explorative Analyse zeigt für unsere Stichprobe tendenziell Unterschiede in der Gewichtung der einzelnen sprachlichen Merkmale. Die Expert*innen der Primarstufe legen eine stärkere Gewichtung auf Verständlichkeit. Die Expert*innen für die Sekundarstufe scheinen demgegenüber sprachliche Merkmale wie die Aussprache höher zu bewerten. An dieser Stelle sei nochmal darauf verwiesen, dass es sich in der vorliegenden Studie um sehr kleine Fallzahlen handelt und Aussagen darüber

nur explorativ zu verstehen sind. Für weiterführenden Studien sollte mit einer größeren Stichprobe eine inferenzstatistische Absicherung erfolgen. Hierbei könnte dann überprüft werden, ob Beurteilende für die Primarstufe die sprachlichen Merkmale in ihrem Gesamturteil anders gewichten als Beurteilende für die Sekundarstufe und somit die Personencharakteristika (Fachwissen, fachdidaktisches Wissen oder beruflicher Hintergrund) die Beurteilungen beeinflussen.

7 Fazit: Limitationen und Forschungsperspektiven

Das DiaCoM-Modell bietet eine theoretische Rahmung, um kognitive Prozesse beim Beurteilen von Lehrpersonen zu untersuchen. Da das Modell bewusst allgemein gehalten wurde, ist es nicht spezifisch auf fremdsprachliche Kontexte angepasst. Ergänzend zum Rahmenmodell von Eckes (2010), welches detailliert mögliche Einflussfaktoren beim Beurteilen von Sprachkompetenzen aufzeigt, liegt der Fokus beim DiaCoM-Rahmenmodell vor allem auf dem Bereich des diagnostischen Denkens. In der vorliegenden Studie wurde eine Erfassung der Diagnose durch ein videobasiertes Beurteilungsinstrument realisiert. Um Rückschlüsse vom beobachtbaren Verhalten auf das diagnostische Denken (s. Abb. 1) erhalten zu können, wurden zwei Urteilsmodi (holistisches Gesamturteil und Globalurteil nach einem analytischen Beurteilungsraster) verglichen. Es konnte gezeigt werden, dass trotz unterschiedlicher Bewertungsmaßstäbe zwischen den einzelnen Beurteilenden die mündlichen Sprachproben der Schüler*innen zuverlässig beurteilt werden konnten. Explorativ wurde außerdem untersucht, ob der berufliche Hintergrund der Beurteilenden (hier Ausbildungstyp) die Integration der einzelnen sprachlichen Merkmale in ein Gesamturteil beeinflussen. Die Unterschiede zwischen Beurteilenden verschiedener Lehramtstypen müsste mit einer größeren Stichprobe statistisch abgesichert werden.

Der Einfluss der Urteilsmodi ist bei der Expertengruppe nur gering (ähnlich wie bei Porsch 2010 für Schreibleistungen). Inwieweit das am Beginn vorgenommene holistische Gesamturteil bei der gesamten Gruppe der Expert*innen einen Einfluss auf die Bewertung der einzelnen sprachlichen Merkmale ausübte, wurde nicht untersucht. Diskutiert werden könnte ebenfalls, inwieweit das wiederholte Zeigen der Videos einen möglichen Einfluss auf anschließende Urteile ausübt (Halo-Effekt).

Im Hinblick auf die verwendete Bewertungsskala muss einschränkend genannt werden, dass sich die Kommunikationsstrategien nicht eindeutig positiv oder negativ polen lassen. Da diese negativ miteinander korrelieren, deutet es darauf hin, dass die Mehrheit der Expert*innen die Kommunikationsstrategien offensichtlich negativ für ihr Gesamturteil auslegten, d. h. im Sinne von Kompensationsstrategien aufgrund mangelnder Sprachkompetenz. Laut Aussagen eines Experten in seinem Stimulated-Recall-Protokoll hätte er sich eine Unterteilung zwischen sprachfördernden Strategien und Kompensationsstrategien bei den analytischen Merkmalen gewünscht, um sprachbegleitende Gesten, die aufzeigen, wie „die Fremdsprache plötzlich lebendig“ wird, positiv beurteilen zu können. Für weitere Erhebungen würde sich eine Unterteilung in kommunikationsfördernde Strategien (z. B. Umschreiben, Intonation oder Nutzen von Pausen) und kompensatorische Strategien (z. B. Sprachwechsel, Lautmalerei) empfehlen.

Die Ergebnisse dieser Studie und vor allem die hohe Interrater-Reliabilität der Expert*innen können bei nachfolgenden Studien genutzt werden. Interessant und z. B. für die Lehramtsausbildung bzw. Fortbildung relevant wären in diesem Zusammenhang empirische Modellierungen mit weiteren Personengruppen (erfahrene Französischlehrkräfte versus Lehramtsstudierende) sowie verschiedenen Lehramts-typen (Grundschul- versus Gymnasialbereich).

Danksagung Wir bedanken uns bei Maya Machunsky für ihre forschungsmethodologische Beratung bei der Durchführung dieser Arbeit.

Förderung Die vorliegende Studie wurde im Rahmen des Forschungs- und Nachwuchskollegs „Diagnostische Kompetenzen von Lehrkräften: Einflüsse, Struktur und Förderung (DiaKom)“ vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg gefördert.

Funding Open Access funding provided by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Bachman, L. F., & Palmer, A. S. (2010). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.
- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12, 239–257.
- Becker-Mrotzek, M. (Hrsg.). (2012). *Mündliche Kommunikation und Gesprächsdidaktik*. Baltmannsweiler: Schneider-Verl. Hohengehren.
- Behr, U., & Kierepka, A. (2002). Einschätzung der Schülerleistung beim Sprechen: dargestellt an englischen und russischen Beispielen. *Fremdsprachen Frühbeginn*, 3, 5–8.
- Birkel, P. (2005). Beurteilungsübereinstimmung bei Mathematikarbeiten? *Journal für Mathematik-Didaktik*, 26, 28–51.
- Bromme, R. (2001). Teacher expertise. In N. J. Smelser, P. B. Baltes & F. E. Weinert (Hrsg.), *International encyclopedia of the social and behavioral sciences* (S. 15459–15465).
- Bromme, R. (2014). *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens*. Münster: Waxmann.
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria: ASCD.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12, 1–15.
- v. Budesu, D. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114, 542–551.
- Burwitz-Melzer, E., & Legutke, M. K. (2004). Die Übergangsproblematik. *Der Fremdsprachliche Unterricht Englisch*, 69, 2–7.
- Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

- Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, 21, 1–44.
- Chuang, Y.-Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *Asian EFL Journal*, 11, 150–173.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290.
- Decke-Cornill, H., & Küster, L. (2014). *Fremdsprachendidaktik: Eine Einführung*. Tübingen: Narr.
- Diehr, B. (2005). TAPS—Testing and Assessment in Primary Schools. Ein Konzept zur Leistungsbewertung im Englischunterricht der Grundschule. *Neusprachliche Mitteilungen aus Wissenschaft und Praxis*, 58, 25–32.
- Diehr, B. (2006a). Fünf Thesen zur Beurteilung der Sprechleistung im ergebnisorientierten Englischunterricht in der Grundschule. In N. Schlüter (Hrsg.), *Fortschritte im frühen Fremdsprachenlernen* (S. 10–18). Berlin: Cornelsen.
- Diehr, B. (2006b). Reden und reden lassen. Leistungsbeurteilung im Fokus des Forschungsprojekts TAPS (Testing and Assessing Spoken English in Primary Schools). *Grundschule*, 9/2006, 36–40.
- Diehr, B., & Frisch, S. (2008). *Mark their words: Sprechleistungen im Englischunterricht der Grundschule fördern und beurteilen*. Braunschweig: Westermann.
- Drese, K. (2004). Lernstandfeststellungen in der Grundschule und ihre Konsequenzen für die Lehrkräfte in der Sekundarstufe I. *Der Fremdsprachliche Unterricht Englisch*, 69, 22–29.
- Drese, K. (2008). *Einschätzung der Sprechleistung von Lernern im Englischunterricht der Grundschule*. Dissertationsschrift. Gießen: Online Universität Bibliothek.
- Eckes, T. (2010). Die Beurteilung sprachlicher Kompetenz auf dem Prüfstand: Fairness in der beurteiler-gestützten Leistungsmessung. In K. Aguado, K. Schramm & V. J. Helmut (Hrsg.), *Fremdsprachliches Handeln beobachten, messen, evaluieren: Neue methodische Ansätze der Kompetenzforschung und der Videographie* (S. 65–97). Frankfurt am Main: Peter Lang.
- Eckes, T. (2015). *Introduction to Many-Facet Rasch measurement: analyzing and evaluating rater-mediated assessments*. Frankfurt a.M.: Peter Lang.
- Eriksson, B. (2012). Leistungsmessung. In M. Becker-Mrotzek (Hrsg.), *Mündliche Kommunikation und Gesprächsdidaktik* (S. 445–457). Baltmannsweiler: Schneider-Verl. Hohengehren.
- Europarat (Hrsg.). (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen lehren beurteilen*. Berlin: Langenscheidt. In Zusammenarbeit mit Sheils, J
- Fäcke, C. (2017). *Fachdidaktik Französisch: Eine Einführung*. Tübingen: Narr Francke Attempto.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Los Angeles: SAGE.
- Fulcher, G. (2015). *Re-examining language testing: a philosophical and social inquiry*. Florence: Taylor and Francis.
- Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research and Development in Education*, 27, 73–82.
- Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 137–152.
- Grotjahn, R., & Kleppin, K. (2008). Bewertung produktiver sprachlicher Leistungen. In B. Tesch, E. Leupold & O. Köller (Hrsg.), *Bildungsstandards Französisch konkret: Sekundarstufe I: Grundlagen Aufgabenbeispiele und Unterrichts Anregungen* (S. 187–204). Berlin: Cornelsen.
- Grotjahn, R., & Kleppin, K. (2017). Kriteriale Evaluation von Schreibkompetenzen. In B. Akukwe, R. Grotjahn & S. Schipolowski (Hrsg.), *Schreibkompetenzen in der Fremdsprache: Aufgabengestaltung, kriterienorientierte Bewertung und Feedback* (S. 117–157). Tübingen: Narr Francke Attempto.
- Gruber, H. (2010). Expertise. In D. H. Rost (Hrsg.), *Handwörterbuch pädagogische Psychologie* (S. 183–189). Weinheim, Basel: Beltz.
- Gwet, K. L. (2014). *Handbook of inter-rater reliability*. Gaithersburg: Advanced Analytics LLC.
- Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, 20, 281–307.
- Henrici, G., Bärenfänger, O., Aguado, K., & Beyer, S. (2003). Mündliche Produktion in der Fremdsprache. *Fremdsprachen Lehren und Lernen*, 30, 3–10.
- Herppich, S., Praetorius, K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, 76, 181–193.
- Hinger, B., & Stadler, W. (2018). *Testen und Bewerten fremdsprachlicher Kompetenzen*. Tübingen: Narr Francke Attempto.
- Hochstetter, J. (2011). *Diagnostische Kompetenz im Englischunterricht der Grundschule: Eine empirische Studie zum Einsatz von Beobachtungsbögen*. Tübingen: Narr.

- Hoge, R., & Coladarsi, T. (1989). Teacher-based judgments of academic achievement: a review of literature. *Review of Educational Research*, 59, 297–313.
- Hu, A., & Leupold, E. (2008). Kompetenzorientierung und Französischunterricht. In B. Tesch, E. Leupold & O. Köller (Hrsg.), *Bildungsstandards Französisch konkret: Sekundarstufe I: Grundlagen Aufgabenbeispiele und Unterrichtsanregungen* (S. 51–84). Berlin: Cornelsen.
- Hutterli, S. (2011). *Beurteilung der Sprechkompetenz in einer Fremdsprache: Theorie und Praxis der Einführung neuer Verfahren in einem schweizerischen Schulsystem (Fallstudie)*. Bern: Peter Lang.
- Institut zur Qualitätsentwicklung im Bildungswesen (2019). Gemeinsame Abituraufgabenpools der Länder – Aufgabensammlung zur Orientierung. Humboldt-Universität zu Berlin. <https://www.iqb.hu-berlin.de/abitur/dokumente/franzoesisch>. Zugegriffen: 20. Mai 2020.
- Knoch, U. (2011a). *Diagnostic writing assessment: the development and validation of a rating scale*. Frankfurt a.M.: Peter Lang GmbH Internationaler Verlag der Wissenschaften.
- Knoch, U. (2011b). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing*, 28, 179–200.
- Kolb, A. (2011). Kontinuität und Brüche: Der Übergang von der Primar- zur Sekundarstufe im Englischunterricht aus der Perspektive der Lehrkräfte. *Zeitschrift für Fremdsprachenforschung*, 22, 145–175.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155–163.
- Krauss, S., & Bruckmaier, G. (2014). Das Experten-Paradigma in der Forschung zum Lehrerberuf. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 241–261). Münster, New York: Waxmann.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods*, 11, 815–852.
- Lenske, G. (2016). *Schülerfeedback in der Grundschule: Untersuchung zur Validität*. Münster, New York: Waxmann.
- Leuders, T., Philipp, K., & Leuders, J. (Hrsg.). (2018). *Diagnostic competence of mathematics teachers: unpacking a complex construct in teacher education and teacher practice*. Cham: Springer.
- Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (Diacom). *Teaching and Teacher Education*, 91, 1–10.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12, 54–71.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Meißner, F.-J. (2005). Zur Einführung in den Themenschwerpunkt – „Neokommunikativer“ Fremdsprachenunterricht. *Fremdsprachen Lehren und Lernen*, 34, 3–14.
- Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes*, 6, 179–189.
- Müller-Hartmann, A., & Ditfurth, S.-v. M. (2011). *Introduction to English language teaching*. Stuttgart: Klett.
- Neumüller, U. E. (2010). Mündlichkeit überprüfen. Mündliche Klassenarbeiten planen, durchführen, bewerten. *Der Fremdsprachliche Unterricht*, 104, 29–35.
- Nieweler, A. (2010). Beurteilung mündlicher Leistungen. In W. Hallet & F. G. Königs (Hrsg.), *Handbuch Fremdsprachendidaktik* (S. 215–219). Seelze-Velber: Klett.
- O'Sullivan, B. (2012). Assessing speaking. In C. A. Coombe, P. Davidson, B. O'Sullivan & S. Stoyloff (Hrsg.), *The Cambridge guide to second language assessment* (S. 234–246). Cambridge: Cambridge University Press.
- Ounis, M. (2017). A comparison between holistic and analytic assessment of speaking. *Journal of Language Teaching and Research*, 8, 679–690.
- Porsch, R. (2010). Die Erprobung eines Kodierschemas zur Messung der Schreibkompetenz. In R. Porsch, B. Tesch & O. Köller (Hrsg.), *Standardbasierte Testentwicklung und Leistungsmessung: Französisch in der Sekundarstufe I* (S. 267–285). Münster: Waxmann.
- Schিপolowski, S., & Böhme, K. (2016). Assessment of writing ability in secondary education: comparison of analytic and holistic scoring systems for use in large-scale assessments. *L1 Educational Studies in Language and Literature*, 16, 1–22.
- Schrader, F.-W. (2009). Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie*, 23, 237–245.
- Shaw, S. D. (2007). Modelling facets of the assessment of Writing within an ESM environment. *Research Notes*, 27, 14–19.
- Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, 19, 85–95.

- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *Journal of Educational Psychology*, 3, 743–762.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Hrsg.), *Planning and task performance in a second language* (S. 239–273). Amsterdam: Benjamins.
- Tesch, B., & Grotjahn, R. (2010). Messung der fremdsprachlichen Sprechkompetenz im Fach Französisch. In R. Porsch, B. Tesch & O. Köller (Hrsg.), *Standardbasierte Testentwicklung und Leistungsmessung: Französisch in der Sekundarstufe I* (S. 177–205). Münster: Waxmann.
- Thiele, S. (2012). *Didaktik der romanischen Sprachen: Praxisorientierte Ansätze für den Französisch-Italienisch- und Spanischunterricht*. Berlin: De Gruyter.
- Thomas, R., Zumbo, B., Kwan, E., & Schweitzer, L. (2014). On Johnson's (2000) relative weights method for assessing variable importance: a reanalysis. *Multivariate Behavioral Research*, 49, 329–338.
- Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: a useful supplement to regression analysis. *Journal of Business and Psychology*, 26, 1–9.
- Vogt, K. (2006). Instrumente zur Leistungsbewertung der mündlichen Sprachproduktion. In N. Schlüter (Hrsg.), *Fortschritte im frühen Fremdsprachenlernen* (S. 19–26). Berlin: Cornelsen.
- Weh, S. A. (2013). *Ausbildung diagnostischer Kompetenzen Konzeption eines fremdsprachendidaktischen Hochschulmoduls zur Diagnostik der mündlichen Sprachkompetenz von Grundschulern: Ein ganzheitlicher Ansatz durch die fallorientierte Arbeit mit Lernerprofilen*. Diss. Heidelberg: Pädagog. Hochschule.
- Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
- Zydati, W. (2005). Leistung feststellen und bewerten.: Im Spiegel der neuesten Bildungspolitischen Entwicklungen. *Praxis Fremdsprachenunterricht*, 5, 3–8.