Fostering and Assessing Pre-Service English Teachers' Oral Teacher Language Competence through an Assessment Rubric and Peer Feedback: An LSP Approach.

RÜTTI-JOY, OLIVIA

BALSTHAL (SO)

2022

DISSERTATION ZUR ERLANGUNG DER DOKTORWÜRDE AN DER PHILOSOPHISCHEN FAKULTÄT DER UNIVERSITÄT FREIBURG (SCHWEIZ)

GENEHMIGT VON DER PHILOSOPHISCHEN FAKULTÄT AUF ANTRAG VON PROF. DR. THOMAS STUDER, DR. CATHERINE DIEDERICH UND PROF. DR. MICHAEL BECKER-MROTZEK

FREIBURG, DEN 11. FEBRUAR 2022

PROF. DR. DOMINIK SCHÖBI, DEKAN

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Institut für Mehrsprachigkeit
Philosophische Fakultät
Universität Freiburg (Switzerland)

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Dissertation eigenständig und ohne unzulässige fremde Hilfe verfasst und noch keiner anderen Fakultät vorgelegt habe.

To Nathaniel, my SMBPPM.

Abstract

Multitudes of factors influence a language learners' success in acquiring an additional language in school. One of those constitutes L2 teachers' language competences in the target language they intend to teach (Vicente, 2012). In addition to high general L2 proficiency, specific, profession-related language competences are receiving increasing interest. For instance, recent theoretical considerations and empirical research suggest that teacher language competence is distinct from general language competence (Cullen, 1998). Indeed, high general and academic language proficiency do not seem to suffice to ensure effective, action-oriented and target-audience appropriate teaching (Bleichenbacher et al., 2017; Bleichenbacher et al., 2014; Burke, 2015; Elder, 2001; Legutke, 2012; Loder-Büchel, 2014). With the rise of competence- and standard-orientation in education, an increasing focus is placed on needs-oriented approaches to identify what L2 teacher language competences are actually required in the real-world classroom. One such needs- and action-oriented attempt constitutes the development of the profession-related language competence profiles (PRLCP) and the corresponding analytical profession-related language competence assessment rubric (PRLC-R) (Kuster et al., 2014). They were devised within a nation-wide Swiss development project and describe the specific language requirements for L2 teachers according to a range of profession-specific communicative skills such as *preparing and conducting lessons* or *assessing, giving feedback and advising* (ibid.). The latter constitutes a central component of fostering learners' L2 skills and is considered a particularly typical, profession-related skill for (L2) teachers (Bleichenbacher et al., 2014). Acquiring the linguistic means of being able to engage in effective L2 feedback practice thus seems of particular relevance for L2 teachers. According to socio-constructivist approaches to L2 education, in which feedback is conceptualised as a multidirectional, collaborative and iterative process, both the feedback provider and recipient take on mutual responsibilities for co-constructing meaning and thus for contributing to successful, dialogic feedback conversations (Carless, 2020a; Carless & Boud, 2018). Accordingly, high language proficiency – particularly of the L2 teacher – are especially relevant to ensure successful reciprocal feedback. Examining how feedback-related teacher language competences can be fostered thus seems to be a necessary step when it comes to exploring the PRLCP in context.

So far, very little accompanying empirical research has been conducted to determine practicality, usability, impact and effectiveness of the PRLCP and PRLC-R in practice. This dissertation examines the practical implementation and systemic relevance of both instruments in two partial studies. By means of a quasi-experimental intervention study of pre-post experimental-control design, partial study 1 investigates how qualitative, language-specific aspects of pre-service English teachers' oral feedbacks provided to lower secondary school students develop under the administration of the PRLC-R in combination with systematic feedback training. The treatment involves the experimental group providing regular feedback on their peers' microteachings based on the PRLC-R including linguistic *and* indigenous criteria (e.g., *addressee-specificity*). The control group identify their own assessment criteria for providing peer feedback. To measure the 48 participating pre-service teachers' oral, profession-related feedback competences in English, an online competence-oriented performance test is used. The pre- and post-tests are based on the PRLCP as the test construct and contain vignette-based test tasks to simulate the target language use domain and elicit relevant oral language performances. The audio-recorded task responses are judged against the PRLC-R criteria by four expert raters. Partial study 2 seeks to answer the overarching research question of how lower secondary school students (i.e. "field experts") perceive and evaluate the linguistic quality and comprehensibility of pre-service English teachers' oral feedbacks. The aim of this sub-study is to explore the perceptions of the end users as stakeholders. The learner judgements are captured through semi-structured, guided interviews. Connections to the PRLCP and the corresponding expert ratings are drawn.

Results of the interrater reliability calculations and rater analyses of partial study 1 show that interrater reliability could not be achieved. In addition, the Multifaceted Rasch Analysis (MFRA) indicates noticeable rater and interaction effects and severe differential rater functioning, suggesting that the PRLC-R criteria are not distinct and reliably applicable throughout. In addition, despite correcting the rater biases and variability through an MFRA, the pre-post analyses show that no treatment effects can be observed in the experimental group. The control group's measured oral, profession-related feedback competences increased by a small, albeit non-significant amount. Findings from partial study 2 indicate that the indigenous criterion *addressee-specificity* may constitute its own independent construct. The insights gained through the intervention study and qualitative interviews serve to identify the potential affordances of the PRLCP. The findings also aid to recognise areas for further development of

the PRLCP, PRLC-R and the indigenous criterion *addressee-specificity*. Implications on the construct of teacher language competence and L2 teacher education are drawn, didactic, methodological and theoretical considerations are presented, avenues for further research are outlined and the need for more accompanying empirical research in development projects are discussed.

Keywords: *Profession-related language competences, teacher language competence, language teacher education, oral feedback competences, L2 speaking performance, performance test, peer-feedback, feedback literacy, learning with rubrics, e-portfolio*

Preamble

Acknowledgements

It takes a village to raise a child. My heartfelt thanks go out to so many people that have supported me throughout my PhD. First, I would like to thank my supervisor Prof. Dr. Thomas Studer and my co-supervisor Dr. Catherine Diederich for being the best role models a novice scholar could wish for. Thank you for your never-ending support, your understanding and your leadership through all the ups and downs. Thomas, thank you for your gentle, empowering and strengthening guidance, for believing in me and valuing all of my ideas (of higher and lower quality) throughout this enriching and powerful learning process. Thank you for showing me the beauty and treasures of quarries when I thought that all they were was but an incoherent and fragmented mess. Catherine, thank you for sitting down with me for hours on end to find possible pathways worth investigating, for opening my eyes to new ways of thinking and learning, for your unconditionally open ear, your presence, your guidance, your highly contagious positive spirit and strength, your endless support and your wonderful friendship.

I would like to pass my special thanks to all the research participants from the PHSG for allowing me to gain insight into your language learning process. Thank you to the students at Sekundarschule Wilen bei Wil for starring in my video-vignettes and to Sandrine Wild for making it possible. Thank you to the students at OZ Buechenwald Gossau for being willing to participate in my interviews and assess pre-service teachers. This study would not have been possible without you.

My warmest thanks go out to Peter Lenz for helping me with the statistical analyses from the humble beginnings to the polished end. Many thanks to the entire IFDS department, with Kristina Ehrsam for your tireless and unconditional support with questionnaire development, data collection, proofreading, rating, all your advice, guidance, friendship, and all the good laughs and deep conversations that kept me afloat. Thank you to Mara De Zanet for your amazing effort during the rating process, for helping me with my data collection, and for giving up so much of your free time to support my undertakings. Many thanks to Wilfrid Kuster, Dr. Lukas Bleichenbacher, Dr. Thomas Roderer, Dr. Robert Hilbe and Dr. Tobias Haug for

supporting me with all my questions, endeavours, difficulties and requests. Many thanks to Dr. Michael Beck and Dr. Robbert Smit for your constructive criticism and your scepticism when it was most necessary, to Dr. Giang Pham for your methodological advice, to Dr. Larissa Schuler for casting a critical eye on my writing, and to Christian Muheim for helping me navigate through the databases and open access peculiarities.

My most heartfelt and deepest gratitude and admiration go out to my amazing husband Nathaniel Frazer Joy. You accompanied me with your unconditional support along the entire way, endured all of my ups and downs, and never ceased to encourage and challenge me. There are no words to describe my gratitude and love. This would not have been possible without you.

Thank you so much to my parents Michèle and Romeo Rütti and my brother Benjamin Rütti, and to my dear friends Dr. Luzia Sauer, Melanie Sampayo, Lara Niederhauser, Dr. Jenny Mendieta Aguilar, and Christina Probst. You always encouraged me, made me feel at home and gave me strength when I felt like I was losing it. I would further like to thank the PHSG (Docs) community for having my back, being fantastic critical friends and moral supporters, and for jumping on board of lots of ideas: Kristina Ehrsam, Dimitra Kolovou, Stefanie Musow, Dr. Lena Hollenstein, Nathalie Rüsch, Johanna Quiring, Dr. Arvid Nagel, Sanja Atanasova, Rahel Schmid, Dagmar Widorski and Dr. Matthias Baumgartner. Thank you to everyone who supported me during the pre-test and took on the role of test supervisors: Stefanie Musow, Stamatina Stamelou-Eccher, Kristina Ehrsam, Mara De Zanet, Catherine Diederich, Thomas Roderer, Séverine Wolf, Sarah Steinmann, Moana Castiello, and Janine Huser. Many thanks to all of the following supporters who helped me conduct the feedback training with the research participants: Sandrine Wild, Tim Clune, Martina Schlauri, and Mara De Zanet. Many thanks to Prof. Dr. Michael Becker-Mrotzek, Prof. Dr. Urška Grum, and Dr. Thomas Eckes, for your advice when I contacted you out of the blue. My warmest thanks go out to Dr. Iris Henseler-Stierlin and Stefanie Graf for their incredible coaching and guidance. Many thanks to Dr. Giuseppe Bertoni, Franziska Kunz, and Lea and Christian Laely for letting me conduct my solo writing weeks at your holiday houses in stunning locations. Without these focused writing weeks, I would not have been able to complete this book within the planned period. Finally, I would like to thank the PHSG and *swissuniversities* for making this project possible. This has been one of the most rewarding yet most challenging experiences that has pushed me to venture into new territories and develop as a novice scholar, teacher, lecturer, and most of all, as a human being.

Notations and Conventions

I wrote this thesis in British English. Quotes and references are kept in their original language with no translations to stay true to the intended denotation. Certain elements, such as research instruments, are kept in German. For the reporting in this dissertation, I decided not to translate these instruments in order not to lose their original meaning and in order to reflect the purpose for which I used them.

This document is based on a template created by Patrik Fuhrer and Pedro De Almeida (Software Engineering Group, Université de Fribourg). The template is open source and available from: https://www.unifr.ch/inf/softeng/en/assets/public/files/thesis_templates/softeng_en_msword.pdf

List of Abbreviations

ACTFL	American Council on the Teaching of Foreign Languages
ACTSPE	Arizona Classroom Teacher Spanish Proficiency Exam
AfL	Assessment for learning
AL	Assessment literacy
ALTE	Association of Language Testers in Europe
AoA	Area of Activity
BA	Bachelor of Arts
BAK	Bundesamt für Kultur (Swiss Federal Office of Culture)
BICS	Basic Interpersonal Communication Skills
BSc	Bachelor of Science
BYOD	Bring your own device
CAE	Cambridge C1 Advanced Exam, i.e. Cambridge English: Advanced
CE	Chief examiner
CEFR	Common European Framework of Reference for Languages
CEFR-CV	Common European Framework of Reference for Languages – Companion Volume
CK	Content knowledge
CLA	Communicative language ability
CLAsS	Classroom Language Assessment Schedule
CLASS	Classroom Assessment Scoring System
CLIL	Content and language integrated learning
CLT	Communicative language teaching
C0	Control group 0
C1	Control group 1
DV	Dependent variable
E	Experimental Group
EDK	Schweizerische Konferenz der kantonalen Erziehungsdirektoren (Swiss Conference of Cantonal Ministers of Education)
EFL	English as a foreign language
ELT	English language teaching
EMI	English-medium instruction

ESP	English for specific purposes
ETP	English-taught programmes
FLT	Foreign language teaching
HarmoS	Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS-Konkordat)
IATEFL	International Association of Teachers of English as a Foreign Language
ICR	Intercoder reliability
IELTS	International English Language Testing System
IFDS	Institut Fachdidaktik Sprachen (Institute for Language Teacher Education)
ILD	International language diploma
IPAF	Individual Presentation Assessment Form
IRR	Interrater reliability
IV	Independent variable
KMK	Kulturministerkonferenz
LAL	Language assessment literacy
LAPP	Language assessments for professional purposes
LMS	Learning management sytem
LPATE	Language Proficiency Assessment for Teachers of English
LPTT	Language Proficiency Test for Teachers
LSP	Language for specific purposes
LT	Language testing
L2	Second/foreign language
MA	Master of Arts
MFRA	Multifaceted Rasch analysis
MLAT	Modern Language Aptitude Test
MSc	Master of Science
MS/V/R	Multi-stage assessments / videos / reflection
NMLPE	National Māori Language Proficiency Examinations
NNS	Non-native speakers
NNSE	Non-native speakers of English
OET	Occupational English Test
OPI	Oral Proficiency Interview
OPIc	Oral Proficiency Interview - Computer
PEAT	Professional English Assessment for Teachers
(P)FB	(Peer) feedback

PCM	Partial credit model
PCK	Pedagogical content knowledge
PHSG	Pädagogische Hochschule St.Gallen
PID	Person ID
PK	Pedagogical knowledge
PLD	Performance level descriptor
PLL	Performance level labels
PMWN	"Plus-minus-what's-next"
PP	Pilot-study participant
PRLCP	Profession-related language competence profiles
PRLC-R	Profession-related language competence assessment rubric
R1-4	Rater 1, Rater 2, Rater 3, Rater 4
RA	Raising awareness
READI	Range, ease of speech, attitude, delivery and interaction
RSE	Retrospective self-evaluation method
RQ	Research question
SCT	Socio-cultural theory
SLA	Second language acquisition
TCS	Transactional communication strategies
TEACH	Taped Evaluation of Assistants' Classroom Handling
TL	Target language
TLA	Teacher language awareness
TEPOLI	Teste de Proficiência Oralem Língua Inglesa
TESOL	Teaching English as a second language
TestDaF	Test Deutsch als Fremdsprache
TLU	Target language use
TOEFL	Test of English as a Foreign Language
UG	Undergraduate
WLE	Weighted likelihood estimation

Table of Contents

List of Figures

List of Tables

1 Introduction

To be able to speak one or several foreign languages is an almost indispensable part of any individual's portfolio in the contemporary globalised world. From this perspective, foreign language (L2) teaching and learning are an area of knowledge and education of uncontested significance. Foreign language teaching and learning research as a subject of scientific inquiry is relatively young (Caspari et al., 2016). At its core lie the multifaceted and dynamic nature of L2 teaching and the foreign language-learning classroom, which render the discipline highly interdisciplinary and complex (Caspari et al., 2016; Königs, 2010). The field of research also distinguishes itself by its pronounced interaction and interdependency of both scientific inquiry and practical application (Reimann, 2020a; Studer, 2019). The collision of these two realms presents both a mutually conducive as well as a mutually dichotomous epistemology where research and practice are in constant tension with one another. The dichotomy manifests itself for instance in the pretense of the universality of scientific inquiry on the one hand and the requirement of action-oriented and situational application on the other (Reimann, 2020a). L2 teaching and learning research aims to connect both poles with reference to the foreign language classroom, employing scientific research quality criteria and methodologies to approach questions of the practical realm and thereby optimising teaching and learning practice in its essence (Reimann, 2020a).

However, the connection of both realms poses its own set of challenges. For instance, development-oriented projects can make large contributions to further optimising L2 (teacher) education, teaching methods and teaching materials. Because development projects are outcome- and product-oriented and highly contextualised in a specific field of application, they are in essence closer to practice than theory on the theory-practice continuum. Third-party stakeholders often represent particular political or policy-driven interests when providing the funding for such projects (e.g., the *Swiss Federal Office of Culture* i.e. Bundesamt für Kultur

BAK[1]), and such development-oriented projects are in most cases tied to tight project timelines. These circumstances tend to result in a lack of time and practical resources for pre- and post-implementation empirical research and the validation of the project products. The value of development-oriented projects for L2 teaching and L2 teacher education is indisputable. However, to ensure that the outcomes of development-oriented projects indeed meet the project desiderata – e.g., to achieve better learning success, or to equip L2 teachers with better tools for assessing their students and providing feedback, etc. – scientific accompanying research is indispensable to gain insights into their practicality, usability, validity, effectiveness and impact and to ensure evidence-based teacher education and practice. It is exactly at the intersection where scientific research and practical application meet where this dissertation is located. The following section introduces the particular kind of intersection this study addresses and outlines the overall research interest.

1.1. Background

Foreign language teaching and learning research as a highly interdisciplinary subject has only recently started to emancipate itself as an independent field of scientific inquiry. A major development that greatly influenced how language teaching and learning is understood was the *communicative turn* (Kommunikative Wende) that occurred in linguistics in the 1970s. This paradigm shift started to place a new and stronger focus on the learner instead of the teacher. Along with repositioning the learner to a more central spot of the learning process, the discourse started to acknowledge the importance of acquiring speaking skills instead of only writing and translation skills – a concept that had dominated the realm previously. The idea of developing speaking skills through the negotiation of meaning in interaction started to take precedence in L2 education (e.g., Zydatiss, 2002). This increasingly dynamic view of knowledge and an increasingly social, communicative view of learning was accompanied by international advancements in Second Language Acquisition (SLA) research (Caspari et al., 2016). The *communicative turn* thus initiated a shift from an input- to an output-orientation in teaching and learning, educational research, and educational politics, which resulted in new conceptualisations of what it means to be able to speak a foreign language. The theoretical concept of *communicative competence* was one of those new developments, and it became one,

[1] The Swiss Federal Office of Culture is a strategic political body that develops and implements the Swiss Confederation's culture policy (for more information see https://www.bak.admin.ch/bak/de/home.html).

if not *the* main focus of research and instruction in L2 teaching and learning (Egli Cuenat, 2014). It introduced a new, competence-oriented approach – one that had also largely been determined by the publication of the Common European Framework of Reference (CEFR; Council of Europe, 2001) as a central European language policy document. The CEFR with its action-oriented approach to *communicative competence* has thereafter become a largely influential document of reference for the development of L2 teaching and learning curricula, language assessments, or teaching materials. It serves as a point of reference to most European language learning curricula as well as all Swiss curricula and the HarmoS educational standards (EDK, 2009, 2012). The CEFR with its competence-orientation and hand-in-hand developments in educational science have, among others, influenced the educational domain as a whole and led to the introduction of national educational standards in European countries (e.g., KMK Bildungsstandards, HarmoS Bildungsstandards, etc.). A radical reorientation of language policy and the teaching and learning discourse followed. While the CEFR and newly established standards made significant contributions to a new (theoretical) understanding of and more ecologically valid approach to communicative language ability and language use, the competence-orientation poses challenges for language teaching and learning research, language teacher education and language teachers alike (Egli Cuenat, 2014). For example, despite the much more elaborated theoretical understanding of *communicative competence*, the specific ways of *how* communicative competence is developed and how appropriate methods can be implemented to achieve this goal remain disputed (Grum, 2012). For language teachers, the catalogue of requirements they need to meet when it comes to structuring and conducting "good foreign language classes" as well as when it comes to their own professional development has become so large and complex that meeting and maintaining the standards has become a major challenge. That a large gap between theory, intended curricular innovation and feasible practice remains prevalent is nothing new. Teacher education curricula and professional development programmes carry the responsibility to convey the latest insights from research, theoretical implications, teaching methods and teaching materials in order to attempt to close this gap – or at least, to narrow it (Egli Cuenat, 2014). In order to meet the requirements of output- and competence-oriented teaching and learning, there is a need for the development and appropriate specification of theoretically sound concepts and accompanying research that shed light on the structure of competences per se, as well as the progression of competence-development (Caspari et al., 2008).

1.2. Research Context

With the increasing competence- and standard-orientation in education, new questions and challenges arise, such as questions regarding the level of proficiency that L2 teachers need to attain in order to be able to successfully pursue their profession. In recent years, there has been a shift away from the notion that high-level language proficiency implies near nativeness or nativeness. For instance, the Companion Volume to the Common European Framework of Reference (CEFR-CV; Council of Europe, 2018, 2020) explicitly rejects the native-speaker-ideal principle. This re-positioning has had cascading effects, especially throughout Europe. In the Swiss educational context, and in alignment with this shift away from the native-speaker-ideal, the minimum standards for Swiss language teachers are not defined by a reference to the native speaker (Egli Cuenat, 2014; Loder-Büchel, 2014). With the public discussion generally orientating itself on the CEFR levels (Council of Europe, 2001), primary and secondary school language teachers are generally expected to attain a CEFR level C1 and C2 in the target language, respectively (EDK, 2017). There are, however, no unified and official standards in Switzerland, and practices are highly heterogeneous across cantons and institutions (Bleichenbacher et al., 2019). Nevertheless, these general guidelines and requirements reflect an implicit underlying assumption that language teachers' L2 proficiency needs to be more highly sophisticated if they teach at a more advanced level (i.e. lower- or upper secondary school) than at a "lower" level (i.e. primary school). While this logic of incremental requirement increase dominates the public discourse, the actual needs of the classroom are largely ignored (cf. Egli Cuenat et al., 2010). This state of affairs has led to considering the need for a specific language teaching profile for L2 teachers in the Swiss public school system to identify the kind of L2 competences that are actually needed in order to successfully teach an L2. In response, a team of Swiss researchers conducted a large-scale needs analysis to do just that. The insights led to the development of the *profession-related language competence profiles* (PRLCP) (Berufsspezifische Sprachkompetenzprofile für Lehrpersonen, die Fremdsprachen unterrichten; Kuster et al., 2014) and a *profession-related language competence assessment rubric* (PRLC-R). While the PRLCP are a collection of can-do descriptors collated in a portfolio that describes the specific language requirements for L2 teachers in Switzerland, the PRLC-R are to serve as a tool for the assessment of teacher language competence. Since their publication, the PRLCP have been widely adopted as a framework of reference across Switzerland (Hunkeler et al., 2009). Indeed, both *swissuniversities* (2015) and the Swiss Conference of

Cantonal Ministers of Education (EDK, 2017) endorse the PRLCP and have issued a catalogue of recommendations that concern the implementation and institutionalisation of the profiles in teacher education curricula across Switzerland. Overall, the PRCLP are to function as a framework of reference for language teacher education and the professional development of language teachers – a framework that is very closely tied to the actual needs of the Swiss L2 classroom. While these recommendations contribute to orientating L2 teacher education towards the real-world needs and thereby narrowing the gap between research and practice, very little empirical research has been conducted to determine the PRLCP's and PRLC-R's practicality, usability, effectiveness and impact. This is partly due to the nature of development projects – an issue that leads to addressing the research aims of the present study in the next section.

1.3. Research Aims

Focusing on developing pre-service teachers' profession-related language competences to meet the practical needs of their future profession is a noble and highly relevant pursuit. The lack of research to empirically investigate the actual implementation and resulting effects of the PRLCP and PRLC-R, and the lack of empirically validated means to reliably, validly and objectively assess profession-related language competences present a necessary (and pressing) avenue for further research. The present study seeks to contribute to meeting the desideratum of conducting further research into (teacher language) competence requirements and competence development (Caspari et al., 2008), and to complement the development- and product-oriented nature of the overall PRLCP project. To reach this aim, an empirical, quasi-experimental investigation of the implementation of PRLC-R and a relevant Area of Activity of the PRLCP is conducted to gain insight into their applicability, effects, usability and systemic relevance. The research questions are explored in two partial studies. Partial study 1 constitutes the main-study, which has a mainly quantitative focus that directly concerns the practical implementation of the PRLCP and PRLC-R in L2 teacher education and a language-testing context. It extracts and focuses on Area of Activity 3 (*assessing, giving feedback and advising*) of the PRLCP and investigates (1) how qualitative, language-specific aspects of pre-service English teachers' oral feedbacks provided to lower secondary school students develop under the administration of the PRLC-R in combination with systematic feedback training. It also investigates (2) the usability and functioning of the PRLC-R from a language-testing and human-rater perspective. Partial study 2 is a small-scale sub-study, which constitutes a

qualitative complement to the overall investigation. It seeks to answer the overarching research question of how lower secondary school students perceive and evaluate the linguistic quality and comprehensibility of pre-service English teachers' oral feedbacks. The below section presents an overview of the overall thesis to guide the reader with reference to how these research questions are addressed.

1.4. Structure of Thesis

After the previous chapter 1 has provided the background for this study, introduced the issues at hand, presented the research desideratum, and described its relevance to L2 teacher education in Switzerland, chapter 2 provides the foundation for the subsequent empirical research study. I present the necessary theoretical background on competence (2.1) and communicative language ability (2.2), which serve as the basis for the subsequent exploration of existing conceptualisations of teacher language competence (2.3). After an excursus on mediation in language teaching to gain a more profound understanding of the central role of a teachers' language skills in the L2 classroom (2.3.2), I provide a detailed description of the PRLCP and the PRLC-R including their contextualisation in Swiss L2 teacher education (2.3.3). This sets the groundwork for zooming into Area of Activity (AoA) 3 of the PRLCP: *assessing, giving feedback and advising,* and more precisely into the communicative language activities and descriptors relating to speaking skills with reference to AoA 3 (i.e. spoken feedback skills). Based on these descriptors, I zoom out again to discuss (spoken) feedback practice and its role in (language) teaching from a theoretical and empirical perspective (2.4). These elaborations serve to outline the working definitions of the central terms I use in the subsequent chapters. Furthermore, considerations of feedback literacy as an emerging reconceptualisation of feedback are presented and connections to teacher language competence drawn. These preliminary considerations on (spoken) feedback lead to a more general discussion on the implications of the nature of speaking on assessment (2.5) to establish a connection to assessing spoken, profession-related language competences. Relevant fundamentals of language testing with a specific focus on testing and evaluating L2 speaking performance conclude the theory chapter. In chapter 3, I consolidate and present existing research findings about ways of developing pre-service teachers' L2 oral (teacher) language competence with particular reference to providing feedback or developing teacher and/or student feedback literacy. I conclude the chapter by outlining the research questions and hypotheses based on the synthesis of the literature (1.5). Building on the previous elaborations, I set out to describe the research

methodology employed in the main-study in chapter 4. There, I outline the research instruments including the PRLC-R and pre- and post-test, study design, intervention and test development (4.3 and 4.4), and a description of the scoring procedures of the test responses (4.5). The subsequent chapter 5 includes all statistical analyses from the main-study. It also provides the inferential results and findings to the main-study's research questions. Chapter 6 discusses the findings with reference to foreign language teaching and learning, L2 teacher education and language testing. It also presents the overall limitations (6.5) of the main-study and ethical considerations (6.6). Implications and didactic consequences of the main-study are presented in chapter 6.7. Chapter 7 outlines the context, research methodology and instruments, participants and research procedure of the qualitative sub-study. The data analyses and results including methodological reflections on reliability issues, overall limitations and ethical concerns are presented in chapter 8. Analogously to the main-study, chapter 9 is dedicated to the discussion of the sub-study including the implications and consequences of the findings for the construct of teacher language competence and research instrument development. Finally, chapter 10 concludes the present dissertation by presenting the overall theoretical, methodological and empirical conclusions and outlining of avenues for further research (10.1).

2 Theoretical Framework

Oral speech production and spoken interaction are unique phenomena of communication with their own distinctive specifics (Luoma, 2009). Speaking a foreign language is also considered one of the most complex skills to master when acquiring a new language (ibid.). Furthermore, spoken language production is one of the most central modes of communication to ensure successful classroom interaction and therefore to promote learning (ibid.). It can thus be argued that L2 teachers' speaking skills are a particularly important component of teacher language competence. In line with the focus of this dissertation, the present chapter outlines the central theoretical concepts that underlie L2 oral speech production in L2 education. I open this chapter by introducing models of communicative competence, which provide the basis for understanding approaches to conceptualising teacher language competence. I then outline the concept of mediation as an influential concept to theorising language-teaching practice. These sections provide the necessary theoretical background for the subsequent discussion of the PRLCP. Accordingly, I then zoom into area of activity 3 (AoA 3) of the PRLCP to discuss in detail the theoretical considerations of feedback, the role of feedback in teacher language competence, and its importance in L2 teaching. After highlighting the significance of oral profession-related language skills in developing teacher and student feedback literacy, I outline general (theoretical) approaches to how oral competence can be assessed. In order to do so, I discuss some fundamentals of language testing and the background to language for specific purpose (LSP) tests as well as communicative language and performance tests. Based on understanding how L2 performance has been understood in language assessment, I then outline ways in which L2 teachers' (oral) profession-related language competences are currently assessed. A comprehensive discussion on scoring L2 performance including the challenges connected to human ratings and a brief summary conclude the theoretical background chapter.

2.1. On Generic Competence

Language teaching and learning as a subject of (scientific) inquiry naturally poses essential questions of how language learning can be monitored and how learning achievements can be assessed. In order to determine a language learner's current level of language ability (language *proficiency* at a given point in time during the language learning process) in a foreign language, or her or his progress with learning a new language over a certain period of time, suitable language assessment instruments are necessary. To be able to develop such instruments, there needs to be a clear understanding of *what* it is that a particular test should measure (the *test construct*), and *how* this area of interest can be measured. Identifying the test construct requires a clear understanding of the constituents it is comprised of. As many language tests seek to measure language *competence* in some form or another, the internal structure of the *competence* at hand needs to be defined. Therefore, identifying and understanding the components of (oral) (teacher) language competence, and how these components manifest themselves in language use, is crucial for developing suitable language assessment instruments. This desideratum presents a challenge that has been addressed numerously through the development of various models to explain (oral) L2 competence. Despite elaborate attempts, the models that exist so far are theoretical constructs that lack empirical evidence; thus, the question regarding the structure of language remains unresolved. As Elena Shohamy (1994) summarises,

> [m]uch of the work in LT [language testing] in the past two decades has been devoted to defining language ability, under the rationale that if there is clear identification of the structure of language, it will be possible to design tests to match such descriptions. (p. 134)

In order to establish models that define language ability, one first needs to have an idea of how *competence* as a generic concept is defined in its basic form. It is helpful to consult adjacent subjects such as educational science – a field of inquiry that is in many ways influential and closely related to language teaching and learning. One of the most influential definitions of *competence* is devised by Franz Weinert (2001) and commonly referred to in educational science. He conceptualises *competence* as a multidimensional construct, namely as

> […] die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die

> Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können. (p. 27-28)

Accordingly, the "Weinert'sche" school of thought conceptualises *competences* as multidimensional dispositions whose individual manifestations are determined by factors such as a person's knowledge, skills, understanding, actions, experience, motivation, etc. (Klieme et al., 2003). Zydatiss (2007) proposes that these dimensions interact with one another dynamically in situational, socio-cultural contexts and during communicative, practical and reflexive actions. Because of the dynamic interaction of a learner's skills, capacities and their motivational and volitional conditions, competences cannot simply be acquired in short episodes of learning (Zydatiss, 2007). The proposed multidimensionality of competences looks as follows:

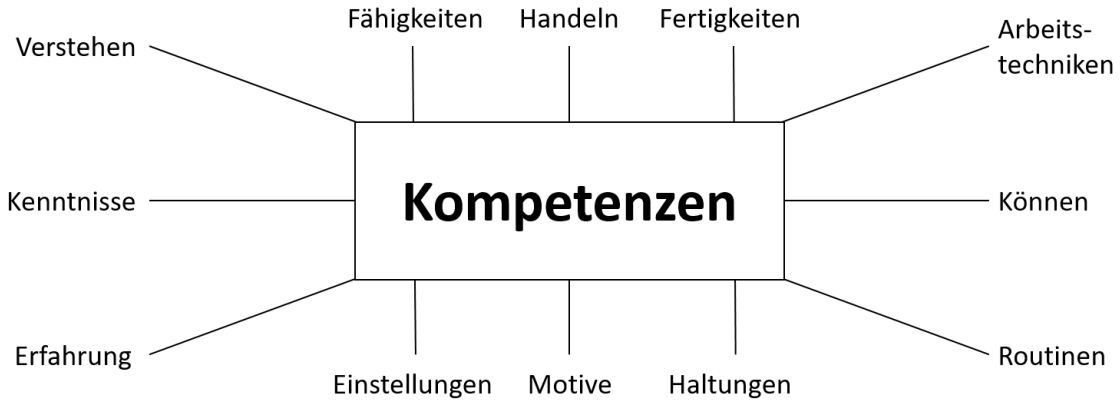


Figure 1 : The multidimensionality of competences (Zydatiss, 2007)

From a higher education perspective, Blömeke, Gustafsson and Shavelson (2015) add to Weinert's and Zydatiss's frameworks by conceptualising competence as a multidimensional construct that stretches "along a continuum from traits that underlie perception, interpretation, and decision-making skills, which in turn give rise to observed behavior in real-world situations" (p. 3). They move away from traditional prevailing dichotomies within the definition – dichotomies that they call "conceptual and statistical controversies" (ibid. p. 11). Instead, they propose an integrated approach with a move towards the complementary interaction of individual components of competence within the specific characteristics and traits of real-life contexts. In sum, and perhaps slightly rudimentary, one can conclude that the above approaches view *competence* as multidimensional, understand that the individual dimensions of competence interact with one another in some integrated and dynamic way, and emphasise that acquiring and measuring competence is complex. In language teaching and learning research,

the understanding of *competence* as a multidimensional construct is reflected in definitions that understand language use as a communicative, interactive, contextualised and goal- and result-oriented (i.e. multidimensional) action (cf. Grum, 2012). However, such multidimensional interpretations are only fairly recent (ibid.). Today's interpretation evolved from its origins by Chomsky and underwent a series of developments. The subsequent section contains an overview of these developments by introducing models of conceptualising language competence. Further, it outlines the current understanding of *language ability* and *communicative competence* to present the central concepts that underlie the construct of teacher language competence on which the overall dissertation builds.

2.2. On Communicative Language Ability

Communicative competence as a concept emerged in applied linguistics in the mid-1970s as a novel approach to language teaching and learning (Luoma, 2009). The previously dominating system relied on largely grammar-focused theories of language competence and focused on analysing language as a system that is disconnected from the language user. In contrast, the communicative competence approach places the language users and language use for communication at its core (Luoma, 2009). This new approach resulted in the *communicative turn*, from whence onwards language teaching and learning became more and more communication- and competence-oriented. The objectives of the L2 classroom increasingly focused on fostering communicative competences as opposed to grammatical knowledge. The communicative turn also radically influenced language testing theories and practices (Grum, 2012). In order to understand this development and the accompanying models of language competence, the origins and development of the concept of *competence* need to be understood. In this subsection, I outline the development of theories on *language ability* and (*communicative*) *competence* by providing a synapse of the chronological progression in educational science, language acquisition and language teaching.

2.2.1. On Language Competence

The term *competence* goes back to Chomsky (1965) who was the first to distinguish between competence and performance. To Chomsky, the difference between *linguistic competence* and *linguistic performance* means that competence refers to a speaker-hearer's subconscious intuitive mental knowledge of language (tacit knowledge of the entire language structure),

while performance manifests itself as a language user's actual use of language in concrete situations:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogenous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. (ibid. p. 3)

Chomsky views linguistic competence superior to linguistic performance, as the latter is always constrained by encoding and decoding processes. From this point of view, performance can never fully denote a language user's *abstract knowledge of language*. Thus, it is not language performance, but linguistic competence, that is prerequisite for a functioning speech community (Chomsky, 1965). While the Chomskian understanding of competence has received a lot of acknowledgement, it has also been criticised for being "monologic", "a priori", and "elementarist" (Habermas, 1970, p. 370; cf. Khan & Taş, 2020) . It is considered reductionist because it claims that language development is based on only the following three criteria: 1) language development occurs in the isolated individual mind, 2) the mind is restrained by inherent cognitive mechanisms, and 3) these cognitive mechanisms rely on a predetermined and limited set of structural rules that lead to (creative) language use (Khan & Taş, 2020). Oppositional views propose, however, that language acquisition and use never happen in a vacuum. Habermas (1970), for example, argues that competence is instead intersubjective (i.e. dialogic), a posteriori (i.e. influenced by experience), and performative (i.e. not tied to a limited set of rules). Hymes (1972) also famously contradicts Chomsky's strict separation of linguistic competence and linguistic performance viewed through a social lens. In particular, he identifies an ambiguity in the way Chomsky conceptualises *linguistic performance*. From Hymes' perspective, the Chomskian view implies two different uses of the term *performance* instead of one as postulated by Chomsky. Consequently, Hymes identifies *performance models* that signify ability as potential, and he distinguishes this *ability for use* from *actual use*, namely instances in which this potential is realised. In this approach, Hymes' (1972) coins for the first time the term *communicative competence*, which in his view encompasses both aspects of knowledge (*knowledge of rules*) and aspects of performance:

> I should take *competence* as the most general term for the capabilities of a person. […] Competence is dependent upon both (tacit) *knowledge* and (ability for) *use. Knowledge*

> is distinct, then, both from competence (as its part) and from systemic possibility (to which its relation is an empirical matter). (Hymes, 1972, p. 282, italics in original)

Chomsky's (1965) focus, in contrast, is a linguistic one with emphasising knowledge of language rather than constructs that underlie performance (McNamara, 1996). Hymes extends Chomsky's concept of competence (1965) by adding "appropriateness of language use" – an essential component that considers that actual language use never happens "in a completely homogeneous speech-community, who knows its language perfectly" (Chomsky, 1965, p. 3). This extended theoretical approach has influenced theorists in applied linguistics (Widdowson, 1990), functional linguistics (Halliday, 1985), language testing research (Bachman & Palmer, 1996), empirical teaching and learning research, and pragmatic educational theory or educational psychology (Grum, 2012; Zydatiss, 2007). They all have in common the underlying assumption that *language ability* can only be observed indirectly through a language user's *performance*. *Performance* results from the underlying abstract knowledge of language – or *competence* – which surfaces through *mediation processes* in concrete communicative situations. The concept of *mediation* was coined by Vygotsky (1978) in his *sociocultural theory* (SCT) where it is placed at the core of knowledge (co)construction, development and human interaction with their environment. SCT understands human mental functioning as a mediated process organised by cultural artefacts, activities, and concepts (Ratner, 2002), which individuals use to regulate, monitor and control their own psychological activity, behaviour and relationship to the physical world (Lantolf et al., 2015). Language is understood as a major mediating tool and powerful cultural artefact that facilitates thought and the construction of ideas (Lantolf et al., 2015) (see chapter 2.3.2). Mapping *mediation* as a concept onto language learning, language constitutes both the auxiliary device and the object of the learning activity itself (Grum, 2012; Widdowson, 1990). Vygotsky's conceptualisation of *mediation* has influenced the development of further conceptualisations of *language ability*, such as for example Widdowson's (1990) approach:

> The language constitutes, in Halliday's terms [(1978)], a meaning *potential*, and this can be manifested through sentences and so internalized. But the potential also needs to be *realised* as use, related to context, made actual, externalized as a purposeful outcome by mediation. It is not enough that the learner knows linguistic resources as an internalized potential, he must also know how to access this knowledge and realise it as a resource. Knowledge of language is a necessary condition for communication but it is not, as Lado [Lado & Fries, 1957] seemed to imply, a sufficient condition. Language is a medium for

> the demonstration of meaning potential but this can only be *realised* by mediation. (p. 123, italics in original)

The original traces of this approach remain prominent until today. For example, the CEFR (Council of Europe, 2001) maintains that competences can never be tested directly:

> All one ever has to go on is a range of performances, from which one seeks to generalise about proficiency. Proficiency can be seen as competence put to use. In this sense, therefore, all tests assess only performance, though one may seek to draw inferences as to the underlying competences from this evidence. (p. 187)

The outlined assumption that *language ability* is located somewhere between the dichotomies of competence and performance, and that it consists of *language knowledge* and *ability for use*, largely influenced subsequent conceptualisations. Two strands of theoretical approaches can thereafter be distinguished. One, unidimensional models conceptualise language ability as one general, overarching type of competence. Such models manifest that language performance can only be assessed globally or holistically as one coherent unit (Grum, 2012). Two, multidimensional models define *language ability* as consisting of several types of competences (or partial competences). Thus, multidimensional conceptualisations build on a divisible competence hypothesis, meaning that a range of partial (sub-)competences interact with one another to construe language ability. Subdividing the construct allows for partial competences to be assessed separately. Both unidimensional and multidimensional conceptualisations of *language ability* are so far merely theoretical assumptions and thus hypothetical models rather than empirically validated frameworks (Grum, 2012). However, there are a number of frameworks that offer differentiated approaches to conceptualising *language ability* that are deemed suitable for language teaching and learning and for language testing research (Grum, 2012). Examples are Canale and Swain's model (1983; 1980) – who were the first to adapt Hymes' (1972) framework to L2 testing –, Bachman's model (1990) and its subsequent extension by Bachman and Palmer (1996), or the CEFR by the Council of Europe (2001). They all outline the constituents of *communicative language ability* and include constituents related to *language knowledge* and to *ability for use* (Douglas, 2010). The following pages provide an introduction to these models.

2.2.2. Early Understandings of Communicative Competence

In their theoretical framework for *communicative competence*, Canale and Swain (1980) suggest that the communicative competence of a language learner is comprised of *language knowledge,* which includes the domains *grammatical competence, sociolinguistic competence* (Hymes, 1972), and *strategic competence*. While *grammatical competence* includes lexical, morphological, syntactical, grammatical, semantical and phonological knowledge, *sociolinguistic competence* subsumes knowledge and appropriate consideration of sociocultural conventions. *Strategic competence,* as addressed in this model for the first time, stands for the possession and activation of "coping strategies" which enable a user to control their communicative behaviour should there be inadequacies in any of the other areas of competence (Canale & Swain, 1980, p. 31). In Canale and Swain's framework, *communicative competence* deliberately excludes *ability for use* and solely refers to language *knowledge*. By rejecting any reference to underlying skills or potential for use because of their resolution that no "theory of human action […] can adequately explicate *ability for use*" (1980), their framework thus fundamentally differs from Hymes' (1972) model. Instead, Canale and Swain (1980) define *ability for use* as *communicative performance*, meaning that performance refers exclusively to a learner's behaviour. They see this *actual use* of the language as *instances of use* where the outlined aspects of language knowledge are demonstrated. In addition, Canale & Swain (1980) emphasise the importance of *strategic competence* to communicative competence, and the need for it to be integrated in "an adequate theory of communicative competence" (Canale, 1983, p. 25). In Canale's subsequent model (1983), the concept of underlying abilities in performance is no longer rejected. Instead, *actual communication* replaces the previous term *performance*, thereby distinguishing its underlying skills and knowledge. Canale specifically emphasises the requirement of such a distinction when modelling *communicative competence* and now aligns more closely with Hymes' (1972) theories: "this notion of skill – how well one can perform knowledge in actual situations – requires a distinction between underlying capacities (competence) and their manifestation in concrete situations (actual communication)" (Canale, 1983, p. 6). In this model, Canale introduces *discourse competence* as a fourth dimension, (formerly *rules of discourse* in Canale and Swain's 1980 framework). *Discourse competence* refers to the "mastery of how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres […]. Unity of a text is achieved through *cohesion* in

form and *coherence* in meaning" (1983, p. 9). The following figure depicts the structure of Canale's extended model:

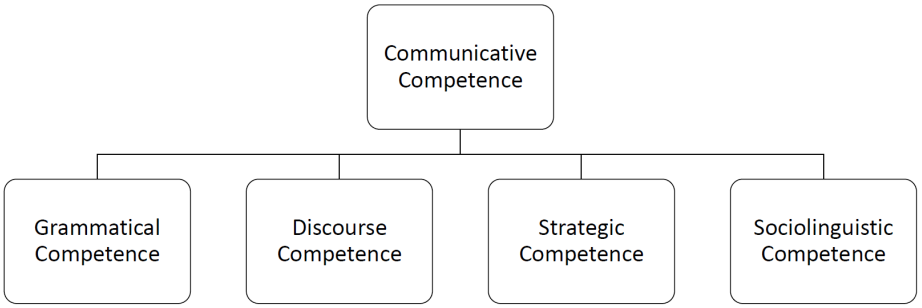


Figure 2 : Canale's (1983) model of communicative competence

Thus, the theoretical frameworks by Canale and Swain (1980) and later by Canale (1983) describe *communicative competence* as a modular construct (Grum, 2012). What these models fail to accomplish, however, is to outline how the individual domains interact with one another.

2.2.3. Communicative Language Ability

Building on Canale and Swain (1980) and Canale (1983), Bachman (1990) devised a more refined model which proposes that *Communicative Language Ability (CLA)* consists of *language competence, strategic competence* and *psychophysiological mechanisms.* In addition, the model includes external factors such as the *context of situation* and *knowledge structures* (knowledge of the world) that influence, if not determine communicative language use. According to Bachman (1990), "[c]ommunicative language ability (CLA) can be described as consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualised communicative language use" (p. 84). The following illustration outlines these components and their relationship to one another:

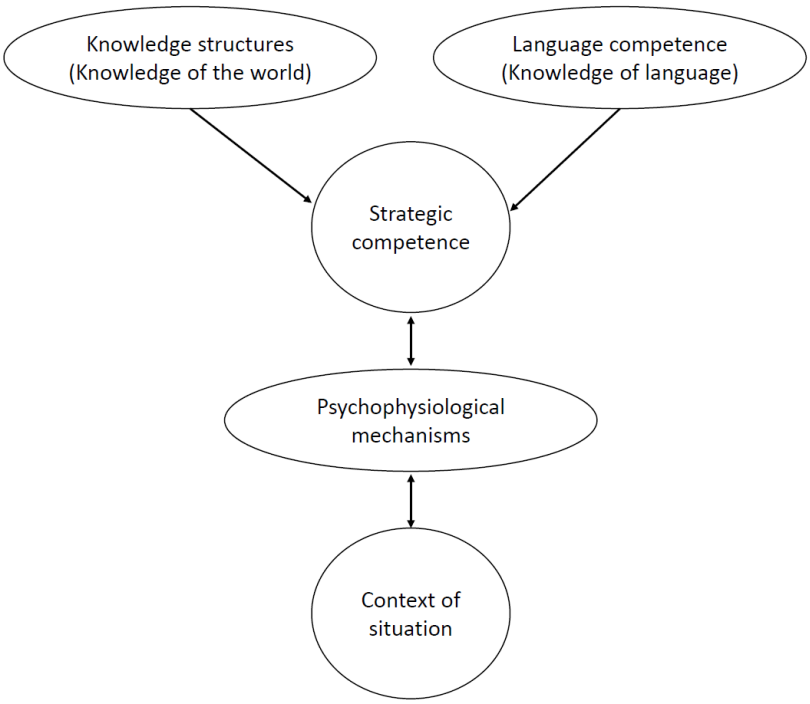


Figure 3 : Components of CLA in communicative language use (Bachman, 1990, p. 85)

Bachman (1990) extends Canale and Swain's (1980) and Canale's (1983) notion of *strategic competence* by placing it at the core of his model. *Strategic competence* is now considered as part of *ability for use* rather than *knowledge.* Indeed, it is not a part of *language competence* but conceptualised as a general ability in its own right that constitutes the nexus between knowledge structures, language competence and individual characteristics of a language user (i.e. psychophysiological mechanisms). These, in turn, are determined by the context of the situation. Thus, Bachman corrects Canale and Swain's (1980) inconsistency in terms of the lack of discussion regarding how the components of communicative competence interact (McNamara, 1996). *Strategic competence* is hence not only a crucial determinant of a language user's overall communicative language ability by enabling them "to make the most effective use of available abilities in carrying out a given task" (Bachman, 1990). Instead, it also plays an important role in the language acquisition process (Grum, 2012). Furthermore, *language competence* as a distinct component of communicative language ability refers to a language user's "control of the rules of usage and use" (Bachman, 1990, p. 105). It consists of the two main categories *organisational competence* and *pragmatic competence*, each of which are again comprised of individual factors, as outlined in the following figure:

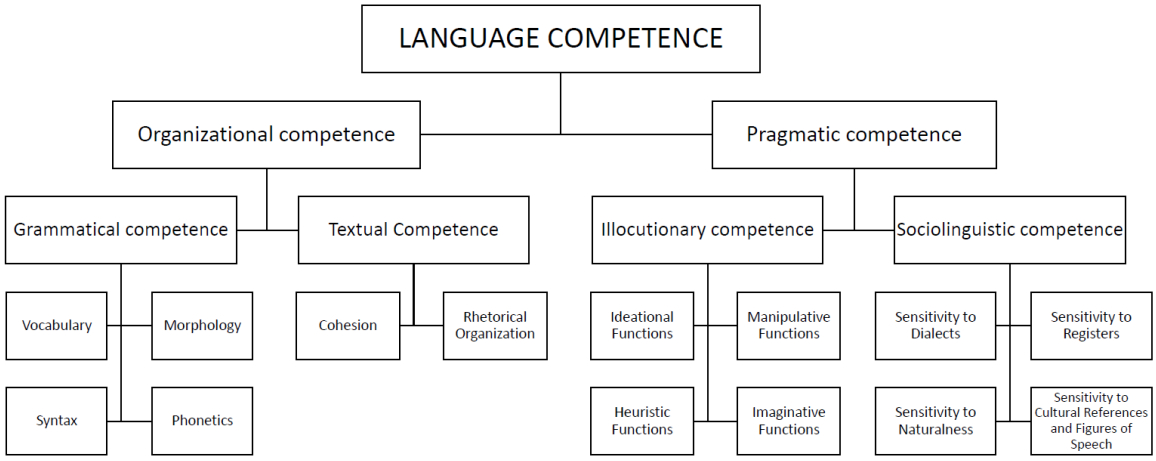


Figure 4 : Components of language competence (cf. Bachman, 1990, p. 87)

Tim McNamara (1996) attributes significance to this model as,

> unlike Canale and Swain, [Bachman] is prepared to open the Pandora's Box of *ability for use* – or at least open it a crack. *Communicative language ability* does include a limited model of underlying capacities in performance, corresponding to at least the cognitive aspects of *ability for use*. (p. 69, italics in original)

In addition, the separation of *strategic competence* from *language competence* is an important step for language testing because it helps to clarify the conceptualisation of language performance by making it more explicit. Indeed, it enables "better investigation of the claims of tests to be assessing communicative language ability" (McNamara, 1996). It is precisely this model that initiated a paradigm shift in language testing to centre on assessing language that has been taught for communicative purposes, i.e. communicative competence (Hoekje, 2016). Bachman and Palmer's (1996) subsequent extensive revision offers a practice-oriented precision of Bachman's 1990 model. Overall, they conceptualise *language ability* as consisting of the three broad dimensions *language knowledge, strategic competence* and *metacognitive strategies.* Their revised model is in this sense revolutionary, as it constitutes a first attempt to approach *ability for use* in relation to affective and volitional factors determined by *attributes of individuals*. Specifically, *ability for use* subsumes cognitive and non-cognitive aspects. Cognitive aspects include *language knowledge* (comprised of *organisational knowledge* and *pragmatic knowledge*), *topical knowledge* (formerly conceptualised by Bachman (1990) as *knowledge structures, knowledge of the world*) and *affective schemata*, all of which can be seen as resources accumulated through previous experiences that a language user draws on (McNamara, 1996). Non-cognitive aspects subsume *strategic competence* as a process dimension that consists of areas of *metacognitive strategy* use (McNamara, 1996). Each of the

metacognitive strategies interact with the components *language knowledge*, *topical knowledge* and *affective schemata*. This interaction becomes apparent in *language use*, which is conceptualised "as the performance of specific situated language use tasks" (Bachman & Palmer, 1996, p. 55). Bachman and Palmer thus move away from the traditional notion of the four skills listening, reading, speaking and writing. Instead, they reconceptualise *language skills* "as the contextualised realisations of the capacity for language use in the performance of specific language use tasks" (ibid. p. 56). By seeing language use as the *performance* of specific and situated activities in which language is used purposefully, the definition of language use becomes more concrete than in the traditional four-skills-concept. In their view, a more useful and pragmatic approach is to reinterpret the traditional "skills" as performed language use that can be described with reference to task characteristics, the areas of language ability engaged, how these aspects are combined, and how they interact in a specific language activity. This *ability-task* concept makes sense when one considers that the entire model is based on language testing research data and is thus strongly orientated towards language assessment. This orientation is in itself revolutionary and becomes apparent in Bachman and Palmer's (1996) description of the two main components that make up *language ability* in relation to language assessment*:* First, both *language knowledge* and *strategic competence* influence language use and assessment. Second, language assessments can now be designed so that these aspects facilitate rather than impede test takers' performance (ibid.). Thus, the new model makes a significant contribution to language testing because it explicitly models the role of non-cognitive (affective) factors that underlie performance. It thereby specifically considers the characteristics of language use in test situations and how they differ from non-test contexts of language use. With this focus Bachman and Palmer (1996) emphasise that assessments of language performance needs to be conducted under the consideration of the language testing conditions and the test task types and functions. In sum, the model reconceptualises *strategic competence* as a set of *metacognitive strategies* and *knowledge structures* as *topical knowledge*, and makes explicit how all conceptualised components of *language ability* are determined by and interact with characteristics of the language use situation, test task setting or test task function. By exploring volitional and affective dimensions of *ability for use*, albeit in a restricted manner, Bachman and Palmer significantly advance Bachman's 1990 model because it accounts for the relevance of these dimensions in language performance.

2.2.4. Communicative Competence in the CEFR

An influential framework that is commonly referred to when it comes to defining and operationalising *communicative competence* is the CEFR (Council of Europe, 2001). This model builds the theoretical framework for the present thesis and is thus described in a more detailed manner. The CEFR provides a multidimensional and extensive competence model that adopts an action-oriented approach including a similar view on the role of *strategic competence* as Bachman's (1990) and Bachman and Palmer's (1996) models. The CEFR is designed and predominantly used in relation to L2 learning, nevertheless it is argued that it is also suitable for all other forms of communication (Coste & Cavalli, 2015). The CEFR's action-oriented approach focuses on the relationship between a language user's implementation of appropriate strategies connected to her or his competences, and on her or his perception or imagination of a communication situation. It also centres on the task a language user seeks to accomplish in a particular context and under specific circumstances. Thus, it views users and learners of a language primarily as *social agents*. Social agents are defined as "members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action" (Council of Europe, 2001, p. 9). According to the CEFR (Council of Europe, 2001), language use and language learning are defined as follows:

> Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of **competences**, both **general** and in particular **communicative language competences**. They draw on the competences at their disposal in various contexts under various **conditions** and under various **constraints** to engage in **language activities** involving **language processes** to produce and/or receive **texts** in relation to **themes** in specific **domains**, activating those **strategies** which seem most appropriate for carrying out the **tasks** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences. (p. 9, emphasis in original)

Individual competences consist of *general competences* that subsume a language learner's knowledge (*savoir*), skills (*savoir-faire*), existential competence (*savoir-être*) and ability to learn (*savoir apprendre*). *Knowledge* stems from experience (empirical knowledge) and formal learning (academic knowledge). *Skills* (or know-how) refer to knowledge that is acquired through experience or repetition. They depend on a learner's ability to complete a process (e.g.,

driving a car or writing a cover letter). While at the outset of learning a skill the steps to carry out a process may need to be consciously broken down into distinct sets of operations, these steps may become automatised with increased experience and practice. *Existential competence* subsumes a learner's individual characteristics, personality traits and attitudes (e.g., cognitive, emotional, volitional, or motivational factors). The underlying assumption is that this type of competence is strongly culture-related and thus is the result of a range of acquisition, acculturation and modification processes. A learner's *ability to learn*

> mobilises existential competence, declarative knowledge and skills, and draws on various types of competence. Ability to learn may also be conceived as 'knowing how, or being disposed, to discover "otherness"' – whether the other is another language, another culture, other people or new areas of knowledge. (ibid. p. 12)

This type of general competence is particularly relevant to language learning as it may draw and be dependent on all the above-mentioned aspects.

These afore-mentioned partial competences are part of *communicative language competence,* which constitutes the second overarching type of competence that is necessary for language learning and use. It comprises *linguistic, sociolinguistic* and *pragmatic* competences, which all require knowledge, skills and know-how. *Linguistic competences* are distinct from their sociolinguistic and pragmatic dimensions and contain lexical, phonological, and syntactical knowledge and skills including other dimensions of language as a system. They also include a language user's cognitive organisation, storage of and access to this knowledge. This accounts for the consideration that this type of knowledge may or may not be conscious and readily accessible, depending on a language user's cultural background and socialisation. *Sociolinguistic competences* subsume the sociocultural conditions of using a language with a sensitivity to social conventions. This sensitivity may be conscious or unconscious, present or absent, and it largely influences all language communication. Finally, *pragmatic competences* contain all aspects that belong to the functional use of a language user's linguistic resources. For example, factors such as knowing and understanding irony, parody and humour, or, among others, different text types and text forms – which all contribute to a language user's discourse abilities – belong to this type of competence. Much like the above-mentioned types of competence, pragmatic competences are constructed through and in interactions within cultural environments and are highly impactful on the success or failure of language communication. A social agent activates her or his *communicative competence* in *language activities* that involve the *reception* (e.g., silent reading)*, production* (e.g., oral presentation)*, interaction* (the

participation of two or more individuals in oral and/or written exchange) with or *mediation* of written or oral texts. The following figure illustrates the outlined CEFR descriptive scheme (Council of Europe, 2018):

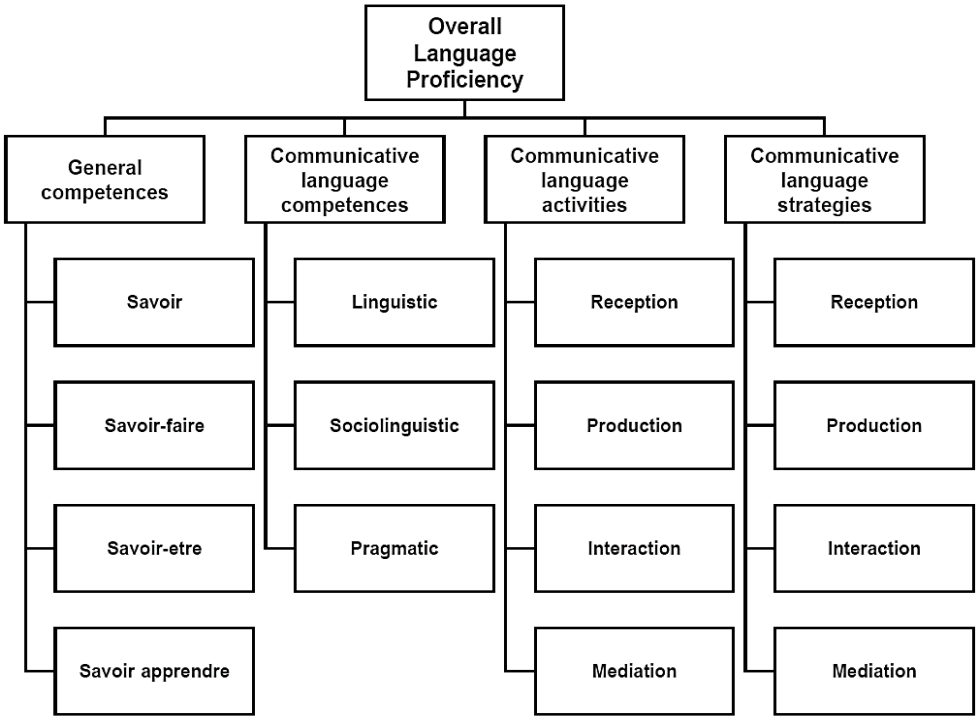


Figure 5 : The CEFR descriptive scheme (Council of Europe, 2018)

The understanding that language ability thus consists of *reception*, *production*, *interaction* and *mediation*, four modes of communication which do not independently coexist but instead overlap, is a move away from the traditional four skills-notion and the view that language ability is comprised of the independent skills *listening*, *reading*, *speaking* and *writing* (Council of Europe, 2020; Lado, 1961). The new approach emphasises the "co-construction of meaning in interaction and constant movement between the individual and social level in language learning, mainly through its vision of the user/learner as a social agent" (Council of Europe, 2020, p. 36; see chapter 2.3.2 for a detailed discussion on mediation).

Turning back to communicative competence in general, the CEFR specifies that any *language activity* is executed in either the *public domain* (e.g., contexts including public services, cultural and leisure activities, etc.), the *personal domain* (e.g., familial and social practices), the *educational domain* (institutional learning and training contexts), or the *occupational domain* (activities related to an individual's vocational occupation). Additionally, competences related to communication and learning are embedded in the performance of *tasks* that are executed in relation to *texts* (written and/or oral). These tasks do not necessarily have to be language tasks.

In order to perform such a task, the social agent employs or devises communication and learning *strategies*. The relationship between such *strategies*, *tasks* and *texts*, then, depends on the nature, context and conditions of the task. Finally, the CEFR defines *communicative competence* as being potentially plurilingual and pluricultural, as its activation draws on "a range of language and cultural resources, which are subject to change and mastered to varying degrees, reflecting the social agent's own experience and involving different languages and language varieties" (Coste & Cavalli, 2015, p. 10). Thus, a language user as a social agent can increase and develop her or his communicative competence through mobilising and implementing the afore-mentioned resources in varied social contexts. The following figure depicts the CEFR model of communicative competence:

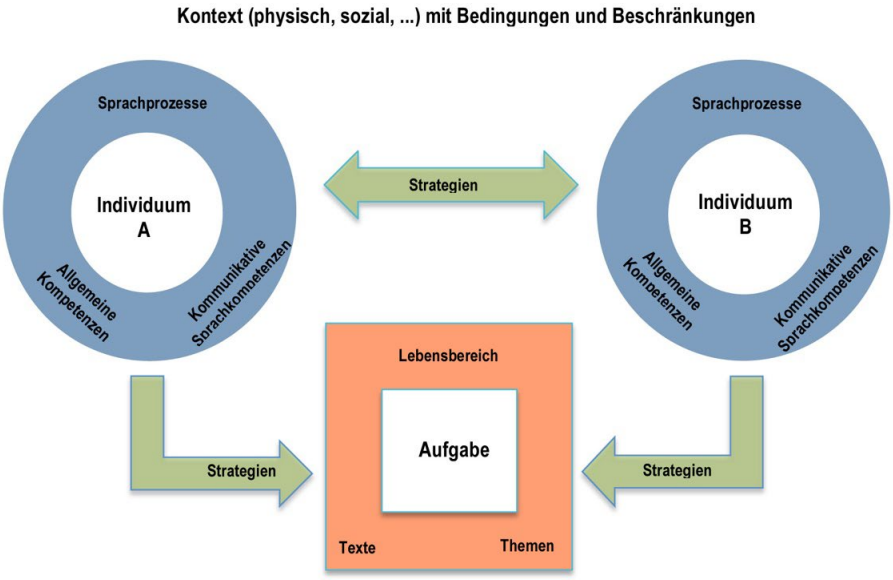


Figure 6 : The CEFR model of communicative competence (EDK, 2012)

In sum, much of the work in language testing devoted to defining *language ability* was conducted under the rationale that a clear identification of the structure of language would enable the development of corresponding and suitable tests (Shohamy, 1994). There is a general consensus that there is a need for an explicit theory of language ability in language assessment or, for that matter, an explicit theory on performance in performance assessment (Douglas, 2010; McNamara, 1996). Models of communicative competence or language ability such as the above recognise and consider how deeply non-cognitive aspects "influence the evolution of the discourse and the interpretation by the participants" (McNamara, 1996). Thus, they enable a more comprehensive understanding of the complexity of language ability and language use. At the same time, such models may pose difficulties in language assessment situations: "we get such a rich picture of the assessment situation that it may be difficult to draw inferences

confidently from it, as it contains too many variables" (ibid. p. 84). Also, the lack of empirical validation means that it remains unclear what the actual components are that make up *communicative language ability,* what exact non-cognitive factors underlie *ability for use* (McNamara, 1996), or how communicative competence is realised in communicative performance (Shohamy, 1994). While questions like these remain unresolved, such theoretical models nevertheless provide a useful and important basis for the development of more contextualised adaptations and the development of language tests (Grum, 2012). The crucial issue to mitigate remains to have a model that conceptualises and includes any relevant underlying capacities of *ability for use* to understand second language performance, but that draws appropriate lines to avoid it becoming unmanageable (ibid.). Because the focal point of the present thesis lies on developing and assessing oral profession-related language competences of L2 teachers, the CEFR framework of communicative competence – and the way it has been devised for adaptations and use – play a central role in conceptualising teacher language competence in this dissertation. Thus, the next section approaches the CEFR model from a contextualised LSP perspective to apprehend the foundations and current understandings of (general and oral) teacher language competence and its constituents in the relevant context.

2.3. Construct of Teacher Language Competence

There is an abundance of literature that discusses what overall competences L2 teachers need so they can teach successfully (Thonhauser, 2019). International publications such as the *European Profile for Language Teacher Education* (Kelly et al., 2004), the *European Portfolio for Student Teachers of Languages*[2] (Newby et al., 2007) *The European Profiling Grid*[3] (North et al., 2013), or *Towards a Common European Framework of Reference for Language Teachers*[4] (ECML, 2017), constitute large-scale attempts of identifying and outlining the professional expertise of language teachers. They provide frameworks of reference for pre-and in-service teachers and teacher educators alike. Language proficiency in the target language constitutes but one of the many required competences an L2 teacher needs to acquire. There is an influential widespread opinion on the particular abilities foreign language teachers should possess. This view postulates "that general English proficiency directly and automatically

[2] https://www.ecml.at/Resources/ECMLresources/tabid/277/ID/51/language/en-GB/Default.aspx (accessed 2.6.2021)
[3] https://egrid.epg-project.eu/ (accessed 2.6.2021)
[4] https://www.ecml.at/Portals/1/5MTP/Bleichenbacher/CEFRLT%20list%20of%20instruments.pdf (accessed 2.6.2021)

qualifies [an L2 teacher] to teach, and that the teacher's general English proficiency directly and automatically improves student learning outcomes" (Freeman, 2017). Statements such as the following by Andrews and McNeill (2005) enjoy ample presence in the literature about L2 teacher competences and L2 teacher effectiveness:

> We have become increasingly convinced that the extent and the adequacy of language teachers' engagement with language content in their professional practice is a crucial variable in determining the quality and effectiveness of any L2 teacher's practice. (p. 159)

Freeman et al. (2009) illustrate this issue by subdividing an L2 teacher's subject matter or content knowledge into two categories: content^{1}, which refers to *knowing language* (i.e. language proficiency), and content^{2}, which refers to *knowing about language* (i.e. knowledge about language and its use). Based on this framework, they explain the native-speaker fallacy:

> Generally speaking, knowledge of and fluency in the target language (*content*^{1}) is taken as a proxy for knowledge about the language (*content*^{2}) […], although the reverse is not the case. Thus, in many settings, when English fluency can be referenced to birth and/or education, which happens in the concept of native speaker […], a teacher candidate who is native is viewed a[s] qualified to teach that language. However, other candidates, who may have in-depth grammatical and meta-linguistic knowledge, but who have not spoken or used the language from birth or perhaps in daily interactions, are seen as less qualified. (ibid. p. 83)

The problem of the idealised native speaker exists on the other end of the spectrum too, for example in the oral proficiency standards for beginning teachers set by the American Council on the Teaching of Foreign Languages (ACTFL). These standards are set based on the flawed assumption that most teacher candidates are native English speakers (Chambless, 2012). However, as Chambless (2012) poignantly illustrates, research has not yet been able to establish a direct connection between a teachers' general proficiency in the target language and effective teaching (ibid. p. 154). Even though this deficit and reductionist ideology has been abandoned *in theory* (Thonhauser, 2019), it seems to remain present in the general public discourse (Freeman et al., 2009) and thereby contributes to undermining the sense of teaching competence for L2 teachers (Freeman, 2017). There is, however, a contrasting view that opposes this proposition and increases in popularity in language teaching research, namely that language teaching requires more than simply high general language competences. That communication in the classroom is different from communication outside of it and thus needs to be

conceptualised differently is nothing new, as for example put forward by Cullen (1998). With reference to *teacher talk*, he called for a reconceptualisation and argued that

> attempts to define communicative talk in the classroom must be based primarily on what is or is not communicative in the context of the classroom itself, rather than on what may or may not be communicative in other contexts; and that the application of criteria of communicativeness solely on the basis of social behaviour which exists in certain contexts outside the classroom could result in an inappropriate and ultimately unattainable model for the majority of language teachers to follow, similar to the earlier preoccupation with teacher talking time. (p. 180-181)

Building on this perspective, one may argue that communication in the classroom demands a specific and unique set of language skills that is different to what is required in other contexts. This view is reflected in the suggestion that high general and academic language proficiency do not suffice to ensure effective, action-oriented and target-audience appropriate teaching that at the same time meets the standards of current language teaching and learning research (Bleichenbacher et al., 2019; Bleichenbacher et al., 2014; Burke, 2015; Elder, 2001; Legutke, 2012; Loder-Büchel, 2014). Burke (2015) emphasises this by stating that high general language ability does not automatically guarantee a teacher's satisfactory performance in the classroom:

> When comparing native speakers to less proficient candidates, Elder (2001) discovered less proficient speakers – potential teachers – could outperform native speakers for certain tasks on the LPTT [(Language Proficiency Test for Teachers)] because of their experiences as second language learners in the language they planned to teach. (p. 4)

The misconception that "native-like" language ability implies teaching proficiency is supported by the fact that international language certificates are often used to certify the language proficiency of (pre-service) teachers and hence serve as gatekeepers to entering the profession. Elder and Kim (2014) accentuate this issue as follows:

> [High currency tests] are likely to underrepresent the teacher proficiency construct. Furthermore, if used on their own, rather than in conjunction with more teacher-specific measures, they are likely to have negative washbacks on the kinds of language teaching and learning undertaken in preparation for performance in a classroom context. (p. 465)

On the one hand, it is to assume that general language certificates do not offer reliable information that allows drawing valid inferences on a language teachers' L2 competences that are prerequisite for developing their ability to teach a language. On the other hand, the construct

of teacher language competence and the interaction of its differential aspects have not yet been fully uncovered. Indeed, there seem to be complex links between a language teachers' language ability and her or his language teaching practice that are in need of further research (Burke, 2015; Legutke, 2012):

> Research […] must be conducted to examine the connection between general language proficiency, academic proficiency, and overall teacher effectiveness. The best speakers may not be the best language teachers. Simply requiring specific levels of general language proficiency and academic proficiency for newly certified teachers will affect who enters the profession, but it is not certain if it will improve world language education. (Burke, 2015, p. 5)

Thus, teacher language proficiency needs to be redefined or further distinguished as an area of competence with a focus on the applicability of language as a skill, or using language as a specific and specialised tool, to fulfil the profession's needs instead of reducing language to a body of content to be learned. So far, concerns about teacher language competence have been prominent in non-English-speaking countries with a focus on *English for Teaching*. Graddol (2006) sees the reason for this preoccupation being the result of globalisation – a development which has led to the assumption that English proficiency takes on a pivotal role in ensuring global economic success and communication. English-speaking countries and English-medium universities are now increasingly showing interest in this issue too, as seen through the mere increase in tests that aim to assess teaching-specific language competences (cf. Elder & Kim, 2014). To develop an LSP test (or any other test, for that matter), a clear definition of the test construct is indispensable. However, there is evidence in the research literature that the expectations, requirements and interpretations of what constitutes teacher language proficiency is highly heterogeneous, that the term is not explicitly defined and that the concept is thus problematic (Hunkeler, 2010; Hunkeler et al., 2009). As Elder and Kim (2014) note,

> [t]he question of what type and level of language proficiency teachers need to teach learners in different contexts, […] remains controversial, with some seeing teacher proficiency as synonymous with native-like competence and others describing teacher language use as a specific purpose domain in which natives and non-natives alike may require training. (p.2)

Several authors have attempted to conceptualise the construct of teacher language proficiency and its partial factors. However, the facts that teaching can occur across a broad range of

disciplines and social and cultural contexts, and that it involves a large variety of tasks, render the deduction of a precise construct definition very complex (Elder & Kim, 2014). The following section provides an overview of what has been done so far by presenting a selection of illustrative approaches to conceptualising the construct of teacher language proficiency. In alignment with *communicative language ability* / *competence*, I seek to take a holistic stance on the entire construct rather than focusing on a teacher's language level at a discrete point in time. I will therefore henceforth refer to the construct as *teacher language competence*.

2.3.1. Conceptualising Teacher Language Competence

A widely known approach to conceptualising teacher language competence is put forward by Catherine Elder (2001). The approach grounds on Shulman's (1987) understanding that a teacher's main responsibility is to "transform the content knowledge he/she possesses into forms that are pedagogically powerful yet adaptive to the variations in ability and background presented by the students" (p. 15). Based on this understanding, she proposes that teacher language competence broadly encompasses "everything that 'normal' language users might be expected to be able to do in the context of both formal and informal communication as well as a range of specialist skills" (p. 152). These *specialist skills* constitute the specificity of teacher language competence and make its difference from general language ability explicit. Elder and Kim (2014) describe *specialist skills* as knowing subject-specific vocabulary and having the discourse competence required to effectively teach in the multifaceted, highly variable and complex environment of the classroom. Additionally, classroom management techniques require specific language and discourse strategies that are unique to the teaching profession. Language teachers find themselves in the special position that in their classes, the foreign language is both the medium and the object of instruction (Elder & Kim, 2014). Thus, in addition to mastering subject-specific terminology, they also "need both metalinguistic knowledge […] and the communicative strategies needed to render this awareness or knowledge comprehensible to and usable by the student" (p. 3, cf. Cullen, 1994). As a final element, Elder (2001) identifies teachers' ability to "provide adequate exposure to rich models of the TL [(target language)] for their students as well as ample opportunities for TL use" (p. 3). She thus supports the argument that teacher language competence involves more than the acquisition of general or academic language skills because of the distinct and unique structure of the (L2) classroom domain. As a result, "native and non-native teachers alike, regardless of their general or academic proficiency level, will require training in appropriate communicative

behaviours for the classroom" (p. 3). While Elder (2001; 2014) succeeds at providing a more comprehensively developed concept of teacher language competence, it remains rather broad.

Another attempt at identifying individual aspects of teacher language competence is proposed by Sabine Doff and Frederike Klippel (2007). They argue that a language teacher must first function as a model for their students, because the quantity and quality of the teacher language they are exposed to markedly influences student success. Second, L2 teachers need to be able to adapt their articulation rate to the language competence level of their students and allow their students ample time for reacting to input or tasks. Apart from adapting their speech rate, Doff and Klippel also mention a teacher's *qualitative adaptation* of their language to what they term *teacher talk* or *teacherese*. Characteristics of this form of *teacher language* are slow speech employing simple structures, i.e. speaking in short sentences of reduced complexity to grant their students access to the foreign language (Wulf, 2001). Other aspects of teacher language competence are related to institutional or domain-specific factors, including didactic questions or didactically motivated framing. In their view, however, the most crucial requirement constitutes a teacher's ability to adequately model natural language use. Doff and Klippel's arguments present some form of dichotomy that is difficult to overcome: on the one hand, a language teacher needs to be a model of good language use that students can emulate and learn from. The language input needs to be complex enough for students to be challenged. On the other hand, the modelled language needs to remain simple enough for students to understand. This is perhaps one of the most striking and challenging aspects of teacher language competence that is implied in Doff and Klippel's argumentation.

In the context of national competence standards and thus teacher competences and teacher professionalisation, Manuela Wipperfürth (2009) argues that the catalogue of teacher competences for teacher education in Germany is inconclusive because it does not comprehensively include specific Foreign Language Teaching (FLT) standards. She puts forth an argument for including more FLT standards based on evidence from FLT research and proposes to incorporate three major additional areas and objectives: *teacher language,* educational objectives of *plurilingualism,* and *intercultural competence*. Based on the FLT research literature, Wipperfürth analysed each of the three areas for the specific demands of L2 teaching and formulated a list of standards consisting of 20 *can-do* descriptors. The following description is restricted to Wipperfürth's analysis on *teacher language*, which she defines as "die Verwendungen von Sprache, die für die Gestaltung und Organisation der Lehr-/Lernprozesse notwendig sind […] sowie Lehrersprache als *comprehensible input*" (p. 13,

emphasis in original). While there is a lack of existing guidelines for what constitutes appropriate or effective teacher talk, there are a series of single findings that Wipperfürth collates which may indicate a move towards some form of clarification of the construct. Primarily, L2 teachers need to be able to provide *comprehensible input* through language that is adapted to the students' level of skill and to their individual needs to neither overtax nor subchallenge them. Like Doff and Klippel (2007), Wipperfürth quotes Wulf (2001) who designates the following language modifications to ensuring *comprehensible input*: reduced articulation rate, pauses, highly clear and precise pronunciation, gesticulation and facial expressions, provision of additional information and didactically motivated questions (e.g., control questions), simplified vocabulary and sentence structures, and simplifications and elaborations such as repetitions or circumlocutions. According to Wipperfürth, L2 teachers need to develop these strategies as well as the ability to employ them in a targeted and appropriate way. Furthermore, L2 teachers need to be able to elicit relevant language productions of students during classroom interactions. To do so, they must develop "[ein] umfassendes und flexibles sprachliches Repertoire, um zum einen vielfältige Schüleräußerungen anregen zu können und deren Interessen mit einzubeziehen; zum anderen um situations- und inhaltsangemessen auf diese reagieren zu können" (p. 15). Based on these elaborations, Wipperfürth suggests five competence standards for L2 teachers. Accordingly, L2 teachers…

- verfügen über eine 'funktional differenzierte, variantenreiche, sichere Kompetenz in der Zielsprache' […], die es ihnen ermöglicht, die Unterrichtsorganisation, allgemeine Gesprächsführung (z.B. auch *small talk*), insbesondere auch Fragen und Feedback *dem Lernstand der Schülerinnen und Schüler angemessen* und zugleich *situations- und inhaltsangemessen* zu gestalten.
- sind motiviert, Gespräche dem Lernstand der Schüler und Schülerinnen angemessen zu führen, deren Sprechanliegen ernst zu nehmen und adäquat auf diese zu reagieren. Sie wissen um die Bedeutung und Formen von Fragen, Feedback und spracherwerbsunterstützender Verwendung der Fremdsprache und können diese effektiv einsetzen.
- verfügen über Strategien bezüglich des angemessenen Wechsels ihrer Rollen als Kommunikationspartner, *instructor* und *facilitator*. Sie können die entsprechenden

Phasen in ihrem Unterricht klar trennen und erlauben so den Schülern und Schülerinnen ein hohes Maß an Redezeit in einem förderlichen Unterrichtsklima.

- können Medien, non-verbale und rituelle Kommunikationsformen zu Hilfe nehmen, um den Schülern und Schülerinnen von Anfang an einen kommunikativ ausgerichteten Fremdsprachenunterricht zu ermöglichen.
- reflektieren den Gebrauch der unter 1 bis 4 genannten Formen in ihrem Unterricht regelmäßig auf der Grundlage einer guten Kenntnis der jeweiligen Schülerbedürfnisse und -fähigkeiten und können sie an diese anpassen. (p. 16, italics in original)

With these standard descriptions, Wipperfürth makes a significant contribution to further specifying the partial competences that make up teacher language competence. As the attempts described above, these descriptions may function well as an overall framework of reference. However, in order to provide a functional basis for the development of potential LSP tests, they are still broad and in need of further specification.

Sokolova (2012) proposes a list of topics foreign language teachers face including the communicative skills and language awareness they require to successfully conduct language lessons. In accordance with previous research on this topic, Sokolova argues that teacher language competence encompasses more "than general language competence due to some elements added and some others deeply interrelated with pedagogical content knowledge and skills" (p. 77). Based on a review of the existing literature she concludes that the specificity of teacher language competence is not the cognitive language load, but the way that load is interrelated with the pedagogical dimension (p. 83). Similarly, it is not grammatical and phonological areas of L2 teachers' language awareness that are distinct from general language, but much more the interrelation of teachers' language awareness with pedagogical content knowledge[5] (PCK) which enables an L2 teacher to teach a foreign language (p. 84). She also states that there is a difference between the language competences needed for inside and outside the classroom, of which both are necessary competences for L2 teachers to successfully function in their profession. By adding out-of-classroom communicative actions to the

[5] Shulman (1987) describes pedagogical content knowledge as representing "the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organised, presented, and adapted to the diverse interests and abilities of learners, and presented for instruction."

repertoire of teacher language competence, Sokolova thus makes a significant contribution to further specifying the construct. Thus, Sokolova defines teacher language competence as:

> *an ability to function successfully both in and out of the classroom achieving professional aims and adapting linguistic and non-linguistic behaviour in accordance with the given context.* It is done through careful selection and application *of language means, general and professional communication skills, classroom strategies, knowledge of language and language teaching rules*. (p. 93, italics in original)

In contrast to the above attempts, Richards, Conway, Roskvist, and Harvey (2013) equate teachers' language competence with teachers' subject knowledge. Indeed, they argue that language teachers' subject knowledge is *the* determining factor that enables them to create a successful language learning experience for their students and to manage key aspects of classroom practice. Extensive subject knowledge also allows L2 teachers to

> effectively adapt or supplement the course book, evaluate the usefulness of the resources […] and make use of authentic materials […] that will prepare and motivate learners to use language outside the classroom. (p. 233)

Furthermore, high language proficiency or subject knowledge is responsible for accurate modelling of target language structures, lexis and pronunciation. They argue that it enables a teacher to provide corrective feedback, accurate explanations that are meaningful to learners and extensive comprehensible input for learners. Finally, it allows teachers to adjust their language according to their learners' L2 proficiency. Subject knowledge in this context is comprised of

> knowledge of second language acquisition theory, pedagogical knowledge, curricular and syllabus knowledge and cultural knowledge, as well as teachers' proficiency in the target language and an awareness of the structure and features of the target language. (p. 232)

Based on Bachman's model for communicative language ability (1990), Richards et al. (2013) conclude that, for an L2 teacher to be proficient in the language they teach, they need to "have an understanding of language systems and be able to use the language for communicative purposes in different situations" (p. 233). While at first sight the equation of language proficiency and subject knowledge seems reductionist and disregarding of language-teacher specific language competences, the skills Richards et al. (2013) list that fall under *subject knowledge* are relatively similar to the more comprehensive conceptualisations presented above. Nevertheless, this view does not account for the multifaceted nature of the classroom

and the specific language skills a language teacher needs to acquire that are different from general language skills. This understanding is a common one that is still relatively prevalent in teaching and learning contexts.

In their paper on rethinking teacher proficiency in the classroom, Freeman et al. (2015) argue that teaching presents a specific context that requires L2 teachers to draw on certain types of communicative abilities. Based on an English for specific purposes (ESP) approach, they attempt to conceptualise teachers' classroom language competence, which they refer to as *English-for-Teaching* (Freeman et al., 2015). Through adopting an ESP approach, they can take into account the implications of situational differences in language for defining the language skills teachers need in the classroom. They thereby consider *English-for-Teaching* to be "both a language and a knowledge construct, which serves to reassemble the dual roles of English – as both the medium and the object of instruction" (p. 4). In Young et al.'s (2014) framework, *English-for-Teaching* is defined as

> [t]he essential English language skills a teacher needs to be able to prepare and enact the lesson in a standardised (usually national) curriculum in English in a way that is recognisable and understandable to other speakers of the language. (p. 5)

The framework identifies the following functional areas that constitute *English-for-Teaching*: Teachers apply language knowledge (1) for managing the classroom, (2) for understanding and communicating subject knowledge, and (3) for assessing and providing feedback. While Freeman et al. (2015), based on Young et al.'s (2014) framework, argue that the *English-for-Teaching* construct "repositions English as a practical communicative tool to carry out certain defined responsibilities within a professional or work context" (p. 6), they do not further specify what this entails exactly. However, on a broader level they argue for the construct addressing and integrating the tension inherent between *global* (i.e. general) and *local* language used (i.e. a teacher's use of language in classroom instruction). Pedagogical knowledge (PK) is infused throughout *English-for-Teaching* as a skill to enact teaching. This infusion is made explicit for example in the way a teacher uses language to manage the classroom, corrects students, provides feedback, etc. Thus, "using the classroom language binds the task with a particular purpose; it is language used to teach" (p. 7). Finally, Freeman et al. (2015) strongly argue for adopting an ESP approach for the assessment of *English-for-Teaching* to increase the validity and level of authenticity of such testing procedures. In a later publication, Freeman (2017) adds that *English-for-Teaching* "defines an asset-, rather than deficit-, based view of the place of

English in ELT [(English language teaching)] teaching knowledge" (p. 50). She emphasises that outmoded ideals of the omnipotent native-speaker are overdue

> to be replaced with the notion that ELT teachers are 'native' to their classrooms. This professional definition of nativeness means that teachers know what they want to do in their teaching; they understand the purposes and uses that English needs to accomplish in their classrooms. What they are seeking is the specific language 'for-teaching' to do so. (p. 50)

Freeman thus broadens the conceptualisation of teacher language competence – even though focused solely on the target language English – to a reconceptualisation of what it means to be *native* in the context of language teaching. This argument stands in an intriguing relationship with the CEFR-CV (Council of Europe, 2018, 2020). While Freeman redefines nativeness in ELT and approaches the idea from a contextual perspective where the classroom constitutes a language teacher's "nativity", the CEFR-CV removes the native speaker ideal – understanding nativeness in its more old-fashioned, ideology-like conceptualisation as "something one is born with (or into)" – from the framework.

In their article, Kissau and Algozzine (2017) investigate L2 teaching effectiveness through the lens of conceptualising different types of content knowledge that are central to language teaching. They use the framework of Ball et al. (2008) which builds on Shulman's (1986) definition that content knowledge encompasses "the amount and organisation of knowledge per se in the mind of the teacher" (p. 9). While Ball et al.'s (2008) framework relates to the field of mathematics instruction, Kissau and Algozzine (2017) extend it to the language teaching domain. The subdomains of the framework include 4 types of content knowledge (CK):

- Common CK: knowledge related to a specific domain that is required in a variety of contexts outside the classroom, here: general knowledge of the target culture as well as general language competence including "knowledge of vocabulary, grammar, expressions, discourse conventions, and the ability to apply this knowledge to communicate ideas effectively" (ibid. p. 117). In other words, common CK refers to a teacher's ability to communicate well in the target language.
- Specialised CK: specialised knowledge that teaching requires to break down and explain the subject matter in such a way that it becomes accessible to students. In language teaching, specialised CK includes an L2 teacher's ability to explain when

and why a learner's language use is grammatically accurate and culturally appropriate in a given context.

- Knowledge of content and students: this type of CK encompasses "a teacher's ability to know what type of student mistakes to anticipate and to explain concepts in a manner that students find accessible, interesting, and engaging" (ibid. p. 117). It also involves a teacher's ability to connect with students.
- Knowledge of content and teaching: knowledge related to maintaining classroom control, applying appropriate teaching strategies, selecting topics, designing materials and sequencing and differentiating instructions to maximise students' learning outcomes. This type of knowledge more or less equals the concept of pedagogical content knowledge (PCK).

As apparent in the framework presented, Kissau and Algozzine (2017) understand CK as an umbrella term that refers to a L2 teacher's different aspects of (profession-related) language competence. CK of language teachers thus involves teachers' ability to

> recognize *when* something is wrong (common content knowledge), […] to understand *why* it is wrong (specialised content knowledge), […] to *predict such errors* among students (knowledge of content and students), and [to] *plan their instruction* to navigate around such difficulties (knowledge of content and teaching) in a way that *meets all learners' needs.* (p. 117, italics mine)

Kissau and Algozzine (2017) to some extent do conceptualise teacher language competence by taking a "content-based" and teacher-centred approach. They recognise that effective teaching requires more than satisfactory general language proficiency (common CK), but other types of CK as well. They thereby state that L2 teachers may display uneven profiles with reference to these types of CK – much in alignment with the contemporary understanding of second language acquisition being a non-linear, uneven process. However, they do not attempt to describe in detail what specific skills the types of CK involves, particularly with reference to language itself.

To conclude, the majority of the above conceptualisations of teacher language competence are relatively vague. Overall, there seems to be consensus that an L2 teacher is both a language user *and* teacher (Dubiner, 2018), which renders L2 teacher's language competence a variation of general language competence with added factors. In addition, most conceptualisations maintain that L2 teacher's language competence encompasses both general *communicative*

language competence, knowledge about language as a system, pedagogical content knowledge, *and* "teacher" language competence including the command of professional terminology and communicative skills. These communicative skills enable language teachers to function successfully in various contexts inside and outside of the classroom. In this context, some authors emphasise the importance of an L2 teacher to be able to adapt their expression to the proficiency level of their students (Doff & Klippel, 2007; Wipperfürth, 2009; Wulf, 2001). What they fail to outline is that such an ability requires at least diagnostic competences, which do not find any direct mention in any of the above models. Finally, there is a mutual call for developing instruments that allow reliable, valid, fair and precise assessment of teacher language competence. However, the construct remains fuzzy and broad and is nowhere near the standard of models of communicative language competence outlined earlier. Additionally, despite the many attempts of assessing teacher language competence, it is fair to postulate that there is for now no single test that can be advertised as testing teacher language competence in its entirety at a specific level. In order to be able to do so, a more precise construct definition is indispensable. Perhaps the new mediation descriptors of the Companion Volume to the CEFR (CEFR-CV, Council of Europe, 2018, 2020) can guide this process and inform the development of a clearer understanding of what teacher language competence constitutes. This option is investigated in the following subchapter.

2.3.2. Mediation in Language Teaching

Mediation occupies a key position in the action-oriented approach to language learning and is considered both a communicative language activity *and* a communicative language strategy (Council of Europe, 2018, 2020). From this point of view, mediation as an all-embracing *nomadic* notion (North & Piccardo, 2016) and occurs whenever there is a bridging and exchanging between different elements and spaces, or wherever the individual and the social interact. North and Piccardo (2016) describe this process as multifaceted and multilayered. In L2 education, mediation is

> concerned with the role of language in processes like the creation of the space and conditions to facilitate communication, understanding and/or learning, the construction and co-construction of new meaning, and/or the conveyance of information. (ibid. p. 20)

Overall, the goal of mediation in language learning is "to reduce the gap between two poles that are distant from or in tension with each other" (Coste & Cavalli, 2015, p. 12). Coste and Cavalli (2015) summarise the concept as follows:

> To mediate is, inter alia, to reformulate, to transcode, to alter linguistically and/or semiotically by rephrasing in the same language, by alternating languages, by switching from oral to written expression or vice versa, by changing genres, by combining text and other modes of representation, or by relying on the resources – both human and technical – present in the immediate environment. (p. 62–63)

In the CEFR-CV (Council of Europe, 2018, 2020), mediation is redefined in accordance with Coste and Cavalli's (2015) definition of "reducing the gap" between social agents:

> In mediation, the user/learner acts as a social agent who creates bridges and helps to construct or convey meaning, sometimes within the same language, sometimes across modalities (e.g. from spoken to signed or vice versa, in cross-modal communication) and sometimes from one language to another (cross-linguistic mediation). The focus is on the role of language in processes like creating the space and conditions for communicating and/or learning, collaborating to construct new meaning, encouraging others to construct or understand new meaning, and passing on new information in an appropriate form. The context can be social, pedagogic, cultural, linguistic or professional. (Council of Europe, 2020, p. 90)

This conceptualisation goes beyond the level of interpersonal exchange and thus includes the goal of gap-reduction between social agents and new concepts. This new definition in the CEFR-CV is much broader than in the CEFR (Council of Europe, 2001), particularly in relation to L2 education (North & Piccardo, 2016), and the new descriptors are applicable not only to the educational domain but also beyond. Coste and Cavalli (2015) propose two forms of mediation; *cognitive mediation* (to provide access to information and knowledge and to competence building), and *relational mediation* (to contribute to interaction, the quality of exchanges and conflict resolution). These two forms are not mutually exclusive. Both essentially involve language as a means of mediation (as defined by the CEFR 2001, but in a considerably expanded form) within social contexts (Coste & Cavalli, 2015). North and Piccardo (2016) isolate four types of mediation, all of which can be allocated to either *cognitive* or *relational mediation* individually or in combination:

- linguistic (inter- or intralinguistic, which is also a form of cultural mediation),

- cultural,
- social (enabling communication between social agents who cannot communicate without external help), and
- pedagogic mediation:
    - facilitating knowledge, encouraging and fostering thinking
    - co-constructing meaning as a member of a community of practice in an educational setting
    - establishing an environment conducive of the above by providing and maintaining space for creativity

It is argued that teaching subsumes all of the above forms of mediation. For ease of understanding, the following figure illustrates the above framework:

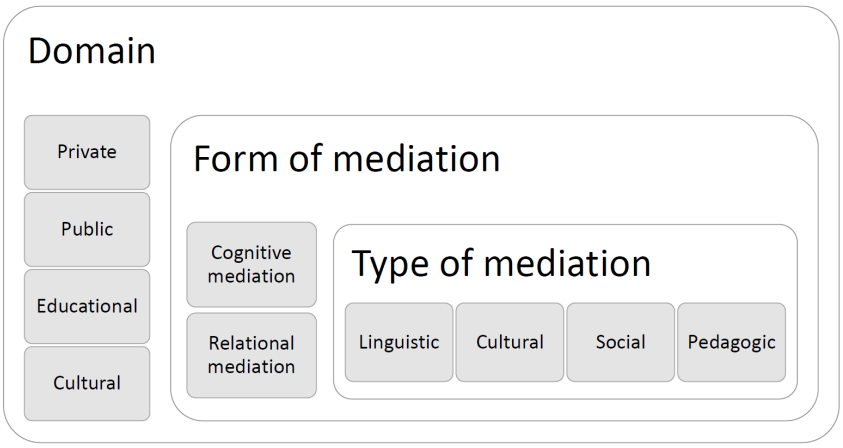


Figure 7 : Conceptualisation of mediation (CEFR-CV, Council of Europe, 2018, 2020)

Based on this framework and with the focus on L2 education, the CEFR-CV contains a vast descriptive system (Council of Europe, 2018, 2020) . For mediation alone, 26 descriptor scales illustrate the construct, thereby distinguishing between mediation *activities* and mediation *strategies*. Mediation activities include:

1) *mediating a text* (transactional language use, e.g., relaying specific information, processing or translating a text, etc.),
2) *mediating concepts* (evaluative, problem-solving language use including both relational and cognitive mediation and the two sub-categories *collaborating in a group* and *leading group work*), and

3) *mediating communication* (creative, interpersonal language use, e.g., facilitating pluricultural space or acting as an intermediary).

While the conceptualisation of mediation in the CEFR (Council of Europe, 2001) encompasses language activities related to linguistic or cross-linguistic mediation (acting as an intermediary to translate or interpret to enable understanding between language users who do not sufficiently understand each other, cf. Council of Europe, 2001, p. 14), the scope of the CEFR-CV mediation scales is much broader. The scales under mediation strategies refer to techniques to clarify meaning and facilitate understanding, and to communicative language or performance strategies that are used in the mediation process (North & Piccardo, 2016) rather than in preparation for it as outlined in the CEFR. The strategies are also applicable the way learners process content. In the CEFR-CV (Council of Europe, 2018, 2020), there are two mediation strategy types: *strategies to explain a new concept* (including *linking to previous knowledge*, *adapting language*, and *breaking down complicated information*) and *strategies to simplify a text* (including amplifying a dense text and streamlining a text). All of the above activities and strategies may combine reception, production and interaction – not only within one, but also across different languages. This alone makes the highly complex and multifaceted nature of the concept explicit. It is thus not too far fetched to assume that conducting (successful) mediation activities likely requires a high degree of cognitive, linguistic and strategic sophistication (ibid.). The below figure provides an overview of the mediation activities and strategies with their respective scales as illustrated in the CEFR-CV (ibid.):

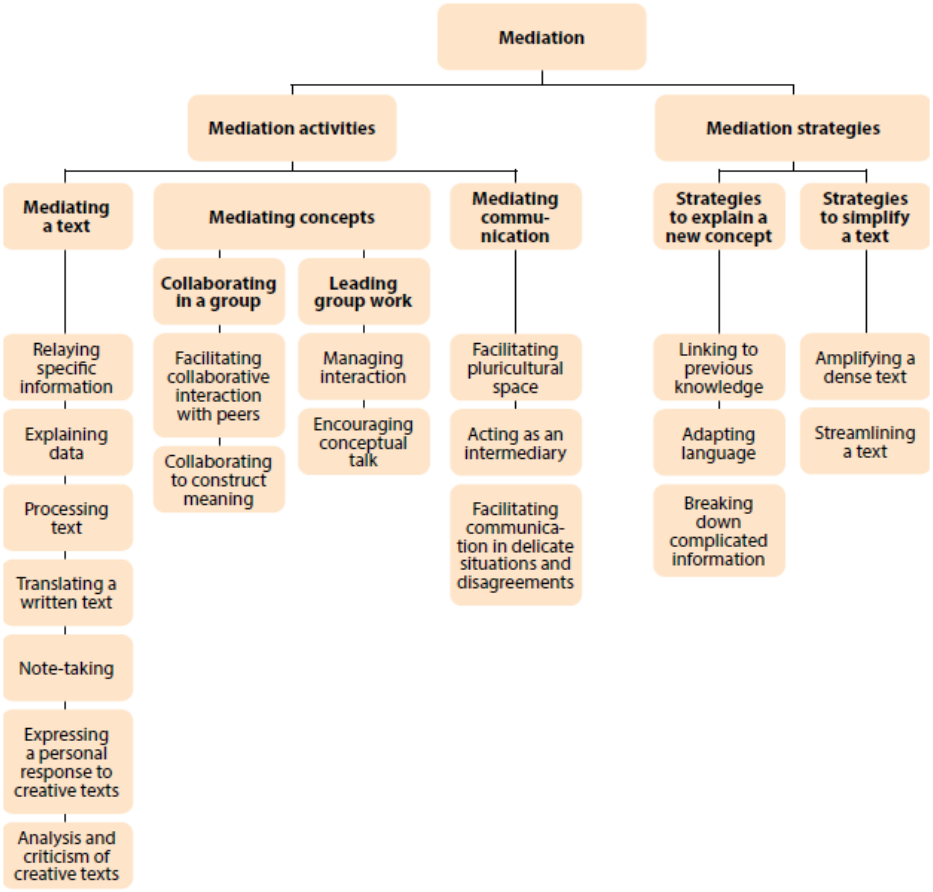


Figure 8 : Mediation activities and strategies (CEFR-CV, Council of Europe, 2018, 2020)

Based on this enhanced conceptualisation of mediation the questions arise 1) what the overall goal of mediation entails, and 2) how mediation is distinguished from other forms of communication. When considering the CEFR-CV (Council of Europe, 2018, 2020), Coste and Cavalli's (2015) and North and Piccardo's (2016) theoretical elaborations, the presumption seems plausible that the uses of mediation are to make information and knowledge accessible and to contribute to developing competence (cognitive mediation). Further assumptions constitute that mediation seeks to afford cognition and activity (Vygotsky, 1986), to facilitate understanding and to assist the active construction of knowledge. These points suggest that mediation contributes to conflict resolution and higher quality interactions (relational mediation). In the educational context, the above activities and strategies can achieve the following goals: to mediate the relationship between the student and something perceived as otherness or *the other*, between the student and another social agent (e.g., another student, a teacher, a foreign culture, etc.), or between a learner's current knowledge and new knowledge (Coste & Cavalli, 2015). Thus, mediation distinguishes itself from other forms of communication because it is tied to the specific purpose of facilitating thought, understanding

and learning. In the educational context and based on the CEFR-CV, teaching can thus only be considered a form of mediation when either texts, concepts or communication are mediated in order to afford activity, cognition or understanding. Such processes involve a mediator who acts as an intermediary between interacting social agents (North & Piccardo, 2016). In doing so, as specified in the CEFR-CV, the mediator tries to reduce the gap between two distant poles (Coste & Cavalli, 2015) and have a "positive influence on aspects of the dynamic relationship between all the participants, including the relationship with him or herself" (p. 107). The poles can be either occupied by individual social agents, social groups, institutions or forms and types of perceived *otherness* (Coste & Cavalli, 2015). Byram (2003) emphasises the unique position of social agents who act as mediators, indicating that this position

> requires them to renounce the categories of thought within which they were socialised and the immediate identity-related solidarities linking them to their community of origin. In order to occupy this third position they must also distance themselves from all the affiliations which are generally involved in communication in the other community with which they are interacting. (p. 96)

Accordingly, it can be argued that mediation takes place in a so-called *third space*, the space in-between, which is an organised, fluid and open space. Coste and Cavalli (2015) specify that *third space*

> is more than a go-between function and more than a kind of filter because, mainly as a result of linguistic variation and reformulation and cultural information and advice, it tends to modify the position of the two poles and bring them closer together through a process of two-way alteration, both sides being affected by change. (p. 29)

I argue that, in the educational domain, this *third space* needs to be seen as a multidimensional space rather than a one-dimensional and linear spectrum to account for the multifaceted, plurilingual and pluricultural, heterogeneous classroom with its diverse social actors. In this context, and according to Coste and Cavalli (2015), a teacher is considered a "professional mediator figure" (p. 29) whose function it is to "provide mediation between pupils and the knowledge, know-how, dispositions and attitudes (savoir-être) that they need or wish to acquire" (ibid. p. 28). North and Piccardo (2016) subsequently specify that teaching *is* mediation, and that successful mediation means "helping learners to appropriate knowledge, but also creating the relationships and conditions to enable them to do so" (p. 15-16). The role of the teacher as a mediating social agent is thus multifaceted and highly complex.

Consequently, it is plausible to assume that mediation in teaching requires a specific set of teacher language competences, which are explicit to the profession and to ensuring effective teaching (or successful mediation, for that matter, as all forms of teaching may be considered a form of mediation according to the elaborations above). In sum, teachers essentially are social agents that mediate in either the cognitive or relational mediation type. Both types

> involve linguistic and semiotic reformulation, a form of language mediation working with terms, texts and discourse genres. [This is because] […] all mediation takes place through discourse: it involves linguistic and discursive (and more broadly semiotic) dimensions, which need to be recognised and effectively used and managed. (Coste & Cavalli, 2015, p. 35)

This mediating action occurs through different mediation activities while employing different mediation strategies. Finally, it occurs with the goal to afford cognition and activity, and to facilitate understanding or the active construction of knowledge. Understandably, however, the mediation scales and descriptors of the CEFR-CV have been criticised for being overly broad (cf. for example Reimann, 2020b). At the same time, they may provide an overall frame of reference, function as a tool to be adapted to specific contexts and act as a framework to develop assessment criteria and (language) teaching and learning materials:

> Die Deskriptoren des Companion eignen sich, in ihrer Ausführlichkeit und häufig weit greifenden Formulierung, […] vor allem auch zur Konzeption von Unterricht sowie zur Bewertung, Klassifikation und (Weiter-) Entwicklung von Unterrichtsmaterialien – und ggf. als Ankerbeispiele für die Entwicklung spezifischer Kriterien und Deskriptoren für einzelne Bildungskontexte. (Reimann, 2020b, p. 17)

That mediation may constitute an important, insightful and potentially highly valuable concept for describing L2 teaching and learning activities as well as outlining potentially worthwhile objectives for L2 learning and assessment seems plausible. Mediation may indeed be seen as a central competence of L2 teachers that require, as mentioned above, specific communicative linguistic, strategic and pragmatic competences. The scales may even be used to complement or even sharpen frameworks that describe teacher language competence, such as those previously outlined or the PRLCP (see chapters 8.2, 8.4, 9.2 and 10.1 for further elaborations). The latter offers a much more precise conceptualisation of teacher language competence than the former; I thus proceed to outline in detail the core of the present dissertation, namely the PRLCP, as a final, more concrete attempt to define and operationalise the construct.

2.3.3. Profession-Related Language Competence Profiles

A series of projects undertaken by a Swiss research team have put forward a further attempt of conceptualising teacher language competence. Switzerland with its specific linguistic landscape and educational context, and its increasingly high language-related expectations of and competence requirements for L2 teachers, continue to challenge L2 teacher education as well as pre-service and in-service teachers alike. With the aim to better align L2 teacher education curricula with the actual language-specific needs of the *real-world* classroom, the *Center for Teachers' Language Competences* (*Fachstelle für berufsspezifische Sprachkompetenzen von Lehrpersonen*[6]) conducted an extensive needs analysis (Long, 2005). The needs analysis was undertaken within the project *Berufsspezifische Sprachkompetenzen von Lehrpersonen* (profession-related language competences of teachers) to precisely identify and operationalise the construct of teacher language competence within the Swiss educational context. The needs analysis involved a systematic review and evaluation of the communicative needs of L2 teachers[7]. This process was realised through the triangulation of a range of different methods and sources of information including the following (Bleichenbacher et al., 2017):

Sources to consult for L2 needs analysis according to Long (2005)	**Sources consulted for PRLCP development**
Reference documents that aid the description of professional tasks and communicative demands of the professional domain	• European Portfolio for Student Teachers of Languages EPOSTL (Newby et al., 2007) • European Profile for Language Teacher Education. A Frame of Reference (Kelly & Grenfell, 2005) • Cadre de référence pour les approches plurielles des langues et des cultures (Candelier et al., 2007) • The INCA Project: Intercultural Competence Assessment (Byram, 2004) • Developing and Assessing Intercultural Communicative Competence. A Guide for Language Teachers and Teacher Educators (Lázár, 2007) • The Common European Framework of Reference CEFR (Council of Europe, 2001) • Profile Deutsch (Glaboniat et al., 2005)

[6] https://www.phsg.ch/de/dienstleistung/fachstellen/fachstelle-fuer-sprachkompetenzen-von-lehrpersonen, accessed on 8.3.2021

[7] See also https://www.phsg.ch/sites/default/files/cms/Forschung/Institute/Institut-Fachdidaktik-Sprachen/201711%20BSSKP%2015-17%20Produktbericht%20Webversion.pdf for full report. Accessed on 5.4.2021

	• Niveau B2 pour le français. Un référentiel (Beacco et al., 2004) • Europäisches Sprachenportfolio (Schneider et al., 2001) • Current Swiss teaching aids of all language and country regions (German, French and English as a foreign language)
Normative documents	Language teaching and learning curricula across the regions of Switzerland: Plan d'Etudes Romand, Lehrplan Passepartout, Lehrplan 21, Entwurf Tessiner Lehrplan
Classroom observations	Observations of teaching / video recordings of classroom situations and interactions (Froidevaux, 2012; Loeliger, 2013; Mettler, 2011; Vicente, 2012)
Consultation with in-service teachers (domain experts) with reference to the communicative demands of their profession	Interviews with in-service teachers (Mettler, 2011), online survey for in-service teachers, "hearings" with in-service teachers across the regions of Switzerland
Consultation with experts	Interviews with L2 teaching and learning experts, linguists, etc.

Table 1 : Sources of information consulted for the PRLCP needs analysis

The analysis of the above sources provided a substantiative basis to deduce communicative language activities which are required of L2 teachers to successfully teach a language (Bleichenbacher et al., 2017). The descriptions of these communicative language activities then served as the basis for the development of a series of tools and guidelines, first and foremost an extensive portfolio of needs-oriented competence profiles (PRLCP[8], Kuster et al., 2014). The PRLCP aim to contextualise the CEFR (Council of Europe, 2001) by continuing its strongly action-oriented approach and mapping it onto the communicative language context of L2 teachers. There is common criticism of the CEFR related to its supposed insufficient consideration of SLA perspectives, insufficiently substantiated intercultural and multi- and plurilingual concepts as well as the lack of substantial considerations of the content quality of speech productions (Grum 2012). However, the CEFR distinguishes itself from other frameworks because the structure of the basic communicative language activities demonstrates the need to distinguish between purely productive and purely interactive (i.e. monological and dialogical) language use and, in yet a separate type of activity, mediation (Grum, 2012). It is precisely because of this distinction that the CEFR offers a suitable framework to use for the delineation of teacher language competence descriptors. Using the CEFR as a framework of

[8]https://www.phsg.ch/sites/default/files/cms/Forschung/Projekte/Berufsspezifische%20Sprachkompetenzprofile%20f%C3%BCr%20Lehrpersonen%20f%C3%BCr%20Fremdsprachen/KP_Sek.I_EN_21.4.15.pdf, accessed on 8.3.2021

reference for profiling profession-related language competences based on the actual professional needs is much in line with Brian North's (2014) intention of use:

> It is really important to emphasise that the CEFR is an instrument to promote profiling and not levelling. [...] All such standards – e.g. to study at university in the language, to teach mathematics in the language, to apply for citizenship – are fairer and more effective when they are based on an appropriate needs profile rather than a blanket ‚level'. This is perhaps the CEFR's main message. (p. 3)

The documents outlining the development and implications of the PRLCP postulate that the profiles describe the specific profession-related language skills that L2 teachers require to fulfil the needs of their vocation. Accordingly, the PRLCP aim to propose a foundation for specifically fostering and evaluating profession-related language competences of L2 teachers, and they are underpinned by the following assumption (Bleichenbacher et al., 2017):

> Fremdsprachenlehrpersonen lehren (und lernen) Sprachen, sie unterrichten mit dem Ziel, die sprachlichen, interkulturellen und methodischen Kompetenzen ihrer Lernenden zu fördern. Dazu bauen sie gelegentlich Aussenkontakte mit zielsprachigen Personen und Institutionen auf. Zudem durchlaufen sie eine spezifische Ausbildung und bilden sich laufend weiter. Sie bewegen sich damit sowohl in schulischen als auch in ausserschulischen Kontexten sowie im Feld der Aus- und Weiterbildung und des lebenslangen Lernens.
>
> Die berufsbezogene Kommunikation von Fremdsprachenlehrpersonen findet somit einerseits **mit Sprachlernenden** statt, für die die Zielsprache in der Regel eine Fremdsprache ist und die sich in der obligatorischen Schule auf einem eher niedrigen Sprachkompetenzniveau befinden, andererseits mit **zielsprachigen Kollegen und Eltern** sowie mit Kollegen und Dozierenden in der **Aus- und Weiterbildung** (hohes bzw. sehr hohes Sprachkompetenzniveau). (p. 12, emphasis in original)

In an action-oriented approach the PRLCP thus describe L2 teacher language competence according to specific communicative language activities and tasks in the communicative contexts outlined above. *Can-do descriptors* specify each communicative language task across different target languages (German, French, Italian and English) and target levels (primary and lower secondary level) in five Areas of Activity (AoA):

> 1: Preparing lessons

2: Conducting lessons

3: Assessing, giving feedback and advising

4: Establishing external contacts

5: Learning and further training

Just like in the CEFR, the *can-do descriptors* express what a learner can do (functional skills) with reference to a specific communicative language task and with reference to a specific competence level (Zydatiss, 2005). In the PRLCP, the *can-do descriptors* of each AoA are organised into communicative language activities and tasks that are conceptualised according to the modalities *production* (speaking and writing), *reception* (reading and listening) and *interaction*. They do, however, exclude *mediation*. The framework was developed before the CEFR-CV (Council of Europe, 2018, 2020) was published with its much broader conceptualisation of *mediation*. With *mediation* being seen as a part of all learning, the mediation descriptors of the CEFR-CV are particularly relevant for the classroom (ibid.). The way the PRLCP were developed allows for post-hoc adaptations and additions to the framework (Bleichenbacher et al., 2017), which, in the case of mediation, would certainly be a valuable further development (see also chapter 2.3.2). Within the currently existing PRLCP framework, the PRLCP adopt a slightly simplified taxonomy, differentiating between the following basic skills only (Bleichenbacher et al., 2017):

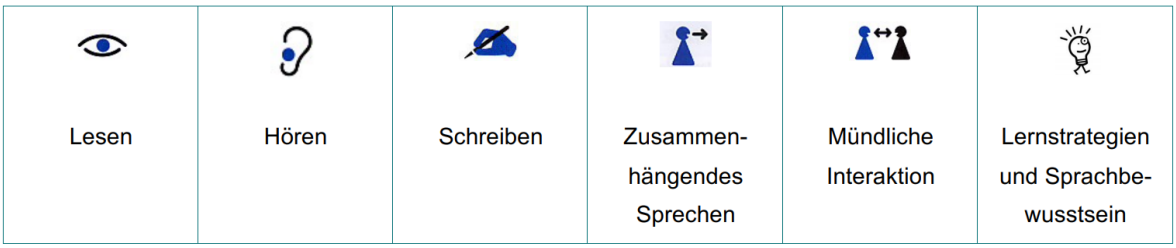


Figure 9 : Taxonomy for classifying communicative language activities in the PRLCP

This implies in some sense a move away from the contemporary Ladoesque (1961) notion of conceptualising *language ability* as being comprised of integrated skills, and a return to the traditional paradigm of understanding the basic skills as separate concepts. Indeed, as outlined in the CEFR-CV, an

> organisation by the four skills does not lend itself to any consideration of purpose or macro-function. The organisation proposed by the CEFR is closer to real-life language

> use, which is grounded in interaction in which meaning is co-constructed. (Council of Europe, 2018, p. 30)

The fact that the PRLCP are built on the four traditional skills and at the same time aim to contextualise the CEFR is somewhat contradictory. On the one hand, the PRLCP claim to be authentic and specifically developed to portray the actual needs of the real-world language classroom. By using the traditional, inert and inflexible model of the four skills, the PRLCP fail to comply with the true nature of the complex reality of language acquisition, communicative language ability and communication. Thus, although the PRLCP indicate an awareness that communicative language ability and profession-related communicative language activities mostly involve a combination of the four skills and thus imply an integrated approach, the regression to the inert tradition remains:

> Gewisse Sprachhandlungen von Lehrpersonen umfassen an sich mehr als einen Fertigkeitsbereich, wie z.B. das Hören und Notizenmachen im Unterricht. In den Sprachkompetenzprofilen werden diese Sprachhandlungen jedoch nur einem Fertigkeitsbereich zugewiesen. (Bleichenbacher et al., 2017, p. 15)

Instead, the PRLCP introduce an additional sixth basic skill, which is mainly applicable to AoA 5 (*Learning and further training*). This sixth skill involves *learning strategies* and *language awareness,* which are particularly relevant for a teacher's personal cultural and linguistic development. By introducing this sixth skill, the PRLCP somewhat compromise for the lack of language acquisition perspective and plurilingual and intercultural concepts the CEFR is criticised for. As summarised by Bleichenbacher et al. (2019), profession-related language competences explicitly involve such skills:

> Neben guten fachlichen Kenntnissen und Fertigkeiten im Bereich der mehrsprachigen, interkulturellen und methodischen Kompetenzen – sowohl für die eigenen Zwecke als auch für den Unterricht – müssen die Lehrpersonen auch über die sprachlichen Mittel verfügen, um im Rahmen geeigneter didaktischer Ansätze die Entwicklung dieser Kompetenzen bei ihren Schülerinnen und Schülern zu fördern. (p. 7)

Thus, the PRLCP attempt to conceptualise teacher language competence at the intersection of communicative language ability and pedagogical knowledge and activities (Candelier et al., 2007; Kelly & Grenfell, 2005; Newby et al., 2007). This combination of activities leads to specific types of communicative language activities or tasks that may or may not be *didactically motivated.* Such tasks may involve, among others, listening to assess a student and give

feedback, or listening and evaluating to be able to adapt one's language to a student's language proficiency. Such communicative activities require basic didactic and pedagogical knowledge and skills in combination with an action- and competence-oriented, multilingual understanding of language learning (Bleichenbacher et al., 2017). In the PRLCP, *didactically motivated communicative language activities* mean the following:

> Als didaktische Sprachhandlung wird in diesem Zusammenhang eine Sprachhandlung bezeichnet, die mit einem bestimmten didaktischen Ziel verbunden ist, also darauf abzielt, etwas zu lehren, einen Lernprozess in Gang zu setzen, einen Prozess zu moderieren, zu stützen usw. Diese Sprachhandlungen erfordern von den Lehrpersonen spezifische Kompetenzen, die sich zum Teil deutlich vom alltagssprachlichen Sprachgebrauch unterscheiden. (ibid. p. 13)

It is precisely this definition that strongly supports the idea of teacher language competence being at times substantially different from general language competence and everyday language use. In subsequent follow-up-projects, and based on this differential understanding within the framework of the PRLCP, instruments for the assessment of profession-related language competences were devised. The specificity of the communicative language needs of teaching requires a different approach to the assessment of the related competences. Precisely because of the multifaceted, multifactorial, highly dynamic and complex nature of teaching and the L2 classroom (Königs, 2010), the rationale required for assessment partly aligns with and partly deviates from the rationale proposed by the CEFR. Indeed, an all-encompassing factor that needs to be considered when evaluating profession-related language skills is a teacher's ability to adapt their expression to the cognitive and linguistic level of their addressees (cf. Doff & Klippel, 2007; Wipperfürth, 2009; Wulf, 2001). This criterion is infused in all devised assessment criteria and thus determines their definition:

> Die **Anpassung des sprachlichen Ausdrucks an die Kompetenzen der Lernenden und anderer Kommunikationspartner** verlangt spezifische Kenntnisse sowie spezifische sprachliche und kommunikative Kompetenzen. An die Qualität der Sprachproduktionen von Fremdsprachenlehrpersonen werden somit spezifische, kontextabhängige Anforderungen gestellt. Die Beurteilung der Qualität der Sprachverwendung von Fremdsprachenlehrpersonen bedingt eine weitergehende **Kontextualisierung der Skalen und Niveaubeschreibungen des GER** bzw. die Entwicklung neuer Skalen und Niveaubeschreibungen. (Bleichenbacher et al., 2017, p. 17, emphasis in original)

Assessing the quality and achieved level or a pre- or in-service teacher's profession-related language competences is thus of complex and highly contextualised nature. The requirements and challenges of this action-oriented approach to (local) language assessment were attempted to be mitigated through the development of assessment instruments based on the above rationale. This included establishing an online platform containing a self- and other-assessment tool for the evaluation of profession-related language competences[9] based on the *can-do descriptors* of the PRLCP. In addition, an analytical profession-related language competence assessment rubric (PRLC-R) was developed. Allen and Tanner (2006) define an assessment rubric, also often referred to as scoring rubric or assessment or scoring grid, as

> a type of matrix that provides scaled levels or achievement or understanding for a set of criteria or dimensions of quality for a given type of performance. (p. 192)

In education, an assessment rubric is comprised of descriptions of levels of performance (Dawson, 2017) that help make learning goals explicit and provide transparent evaluation criteria for learners and teachers alike (Brookhart & Chen, 2015). Rubrics also play an important role in language testing where the scales and performance level descriptors (PLDs) are used (mostly) for the analytical assessment of language performance (see chapter 2.4.3). Hence, aside from providing a new framework for the formative and summative assessment of teacher language competence, the purpose of the PRLCP is to make the linguistic requirements of pre- and in-service teachers more transparent, support curriculum design and act as an aid for the adaptation of current evaluation and assessment formats to the specific contextual needs.

As indicated above, the materials used for the development of the PRLCP are restricted to information from the literature, expert opinions and frameworks of reference. Just like the CEFR, the PRLCP lack empirical evidence collected from language learners. The profiles nevertheless constitute a significant step towards a clearer and more elaborated understanding of teacher language competence. This significant contribution is recognised by both the EDK (2017) and *swissuniversities* (2015) who endorse the integration of the PRLCP in practice, particularly in L2 teacher education. Both recommend the implementation of the profiles and corresponding tools to better align L2 teacher education curricula with the current needs of the teaching profession and to convey, foster and assess the competences necessary for current and high-quality L2 teaching practice. In particular, *swissuniversities* (2015) call for the active use and integration of the PRLCP as a framework of reference within L2 teacher education and

[9] https://profils-langues.ch/, accessed on 8.3.2021

professional development in Switzerland (§1). This integration should focus on developing (§4) and evaluating teachers' profession-related language competences, and on evaluating and improving L2 teacher education curricula by aligning them to the PRLCP. Furthermore, the PRLCP should complement international language diplomas (ILD, §2) and serve as a means to evaluate the language-related requirements placed on pre-service language teachers in L2 teacher education (§3). By following these recommendations, the PRLCP are expected to contribute to the growing competence-orientation in teacher education and professional development (§5). Finally, *swissuniversities* call for continuing developments of tools related to the PRLCP (§6). For example, supplementary materials should be devised to help teacher educators and in-service teachers with the application of the PRLCP in their teaching. Additionally, materials for the evaluation and certification of profession-related language competences should be developed and provided to the cantons and universities of teacher education across Switzerland. In the subsequent 2017 recommendations issued by the EDK (2017) for the advancement of the *Sprachenstrategie*[10] (EDK, 2004, 2014, 2017) and the betterment of the conditions of L2 education in Switzerland, the *swissuniversities* endorsement of the PRLCP is consolidated:

> Den Kantonen und ihren Bildungsinstitutionen wird empfohlen, die Entwicklung und Erhaltung der sprachlichen und didaktischen Kompetenzen der Lehrpersonen zu fördern, […] indem sie die Umsetzung der Empfehlungen der Kammer der Pädagogischen Hochschulen (PH) von *swissuniversities* bezüglich der Einführung der Berufsspezifischen Sprachkompetenzprofile für Lehrpersonen für Fremdsprachen in den Bildungsinstitutionen unterstützen. Es wird empfohlen, möglichst früh während der Ausbildung ein allgemeines Niveau B2 (Primarstufe) bzw. C1 (Sekundarstufe) zu erreichen. Am Ende der Ausbildung soll ein höheres berufsspezifisches Niveau als das allgemeine Niveau B2 bzw. C1 erreicht werden. (p. 4, italics mine)

Thus, the PRLCP have gained prominent status among official stakeholders in Switzerland. The profession-related language competences as described in the PRLCP are now recognised as competences to be achieved by the time pre-service teachers graduate and enter their profession. By implementing the PRLCP and the PRLC-R in the curriculum, the aspired

[10] The (Sprachen)*Strategie der EDK und Arbeitsplan für die gesamtschweizerische Koordination* is a political document that outlines perspectives of improving and homogenising the way foreign languages are learned and taught across Switzerland. This includes harmonising L2 teacher education and aligning the curricula with L2 teaching practice in the actual classroom. See https://www.edk.ch/de/themen/transversal/sprachen-und-austausch for more information (accessed 03.05.2021).

implications should serve to harmonise the competence standards for L2 teachers across Switzerland. Despite the lack of reference to empirical language learner data and little to no empirical evidence of effects connected to their application in the field (see chapter 1), the PRLCP are applied at several universities of L2 teacher education in Switzerland (Fachstelle für Sprachkompetenzen von Lehrpersonen, 2021). The PRLCP represent the most differentiated, elaborated, concrete and evidence-based attempt of conceptualising the construct of teacher language competence that I could locate in the literature. Considering the impact factor of the PRLCP especially, more empirical research is indispensable. The aim of this dissertation is to contribute to this need. Specifically, I aim to focus on those PRLCP descriptors that outline the oral, profession-related language skills of L2 teachers, in particular the descriptors of AoA 3 (*assessing, giving feedback and advising*). In the following section I focus on the central role of L2 teachers' oral profession-related language skills by zooming in on the PRLCP and amplifying AoA 3 by outlining theoretical elaborations underpinning (oral) feedback in the L2 classroom.

2.4. Feedback in Teaching

A large body of research manifests that feedback constitutes an integral part of teaching. Feedback has been deemed one of the most powerful interventions for ensuring student achievement (Hattie, 2009; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Indeed, research suggests that providing students with meaningful feedback is essential if their learning is to be positively impacted and enhanced. It has been shown, for example, that meaningful feedback benefits learners in empowering them to assume ownership of their own learning goals and take control of their learning achievements (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Rasi & Vuojärvi, 2018; van der Kleij et al., 2015). Moreover, giving constructive feedback enables teachers to "step into a student's learning process and potentially change the direction of the process or deepen it" (Rasi & Vuojärvi, 2018, p. 293). Paired with further benefits such as raising awareness of learning goals, making assessment criteria transparent, and providing students with an opportunity to monitor their progress and achievements, feedback can lead to better learning outcomes. If feedback is well-timed, detailed, specific and positive – says the research literature – emotions such as joy, pride and excitement can be triggered that may further empower learners in their learning process (Rowe et al., 2014). With reference to L2 learning, feedback can also act as a valuable source of (comprehensible) input. Andrews (2003) explains this as follows:

> Although [language learners] may encounter L2 input direct from sources such as the textbook […] and other students […], much of the input learners are exposed to involves the teacher. The teacher may be the producer of such input: with the specific intention to induce learning, as in, for example, the presentation of new language; or less deliberately, through any communicative use the teacher makes of L2 in the classroom, such as for classroom management. […] When encountering language produced by the learners, orally or in writing, the teacher has a range of options for handling that output, but very often teacher feedback will provide an additional source of input for learning (for the class or for the individual learner) as the student's original output is modified by the teacher. (p. 90)

It is thus to assume that teachers' elaborated feedback skills is not only of particular relevance to general but also L2 teaching. In other words, one may argue that AoA 3 (*assessing, giving feedback and advising*) of the PRLCP is relevant to L2 teaching and student success. This assumption is supported by the authors of the PRLCP who consider the competences outlined in AoA 3 as highly L2-teacher-profession-specific (Bleichenbacher et al., 2014c). The present chapter discusses the theoretical background of feedback and presents current re-conceptualisations and research findings as evidenced in the literature. Because many of the concepts and findings from the general feedback literature are transferrable to L2 education, the following section weaves in and out of the context of general and L2 teaching. Consequently, the goal of this chapter is to provide a basis to contextualise feedback as a concept in a linguistic and L2-teaching-related, action-oriented frame and hence to describe how feedback, both conceptually and practically, is to be understood in this dissertation.

2.4.1. Terminology and Definition

Although there seems to be a relatively widespread consensus when it comes to the definition of feedback in pedagogical and psychological contexts, a closer investigation of the terminology in the feedback literature suggests that its actual understanding is largely heterogeneous (Müller & Ditton, 2014a). Indeed, the term appears to be comprised of a variety of nuances in meaning and functions, which depend on the respective theoretical or subject-specific approach (Müller & Ditton, 2014a). The root origins of feedback trace back to cybernetics, systems theory and the regulation of machines, organisms and organisations (Müller & Ditton, 2014a; Narciss, 2006). From the mechanistic perspective of cybernetics, the

regulation of dynamic systems implies the continuous registration and re-registration of a control variable by a control device. This process allows for the comparison of the obtained information with a reference variable (Müller & Ditton, 2014a; Narciss, 2006). If the outcome of this comparison indicates a discrepancy between the control and reference variable, the regulation process allows the control variable to be manipulated in order more closely align it with the reference variable. This is commonly referred to as a regulatory circuit or control loop (Narciss, 2006). This cycle continues until the control variable and the reference variable align, at which point the regulation process is complete (Müller & Ditton, 2014a). Ramaparasad's work (1983) builds on this mechanistic understanding of "feedback" by adding an intentional component to it. Accordingly, what constitutes feedback is the fact that the information on "the gap between the actual level and the reference level" is used with the explicit intention to minimise that gap (p. 4). From this point of view, the main purpose of feedback is to intentionally minimise the discrepancy between a current and future state of affairs through altering a specific aspect (Kluger & DeNisi, 1996; Merriam-Webster.com, 2021; Müller & Ditton, 2014a; Ramaprasad, 1983). Hence, it is the effect rather than the content of feedback that allows information to qualify as feedback. In other words, only if information is used to alter the gap between the reference level and current level can it be considered feedback (Ruiz-Primo & Brookhart, 2018). Fast-forward thirty years and peering into second language acquisition (SLA) as a subject of scientific inquiry, feedback retains some of its original meanings and functions in some form, but also changes greatly. In SLA, common terms that are applied in relation to feedback are *positive evidence*, *negative evidence*, *feedback*, and *error correction* (Leeman, 2007). *Evidence* generally refers to information about whether the application of certain structures in the L2 are appropriate and permitted. This information may be provided either before or after a learner produces language. While *positive evidence* indicates that the structures used are permissible in the L2, *negative evidence* points out the opposite (Leeman, 2007). *Feedback*, in contrast, more closely resembles its root origins and refers to information provided to a learner on their language productions, learning processes and achievements. In SLA, feedback may include "information regarding the accuracy, communicative success, or content of learner utterances or discourse, regardless of how the learner interprets and responds to such information" (ibid. p. 113). This deficient interpretation will be addressed further below. Finally, *error correction* is a pedagogical activity that involves offering a learner feedback on their errors. With SLA opening up its focus from theory-building to developing a knowledge-base on effective pedagogy, the conceptualisation of feedback in

SLA has started to include the large body of research on feedback and general learning processes that was previously considered peripheral (Leeman, 2007). However, these newly broadened theoretical positions are still evolving and there is not enough empirical evidence to make any conclusive inferences on the effect of (the various types and forms of) feedback on L2 acquisition (Leeman, 2007). Regardless, Leeman (2007) draws the following tentative implications for L2 theory and pedagogy:

> Like L2 instruction generally, feedback should respond to learners' needs, and therefore the most effective type of feedback will depend to some extent on the source of the L2 error (e.g., non-target L2 linguistic competence, insufficient attention to form, inaccurate declarative knowledge, overgeneralized application of rules). Of course it isn't realistic to think that instructors can make such assessments spontaneously, in the midst of interaction with students, especially given our current state of knowledge on SLA and L2 performance. However, in addition to whatever benefits feedback can provide for the learner, certain types of feedback also have the potential to give instructors a better sense both of the source of the problem and of the learner's current L2 knowledge. (p. 131)

These few introductory remarks on feedback illustrate that it depends largely on the discipline and theoretical framework at hand what exactly feedback implies in practice. They also illustrate that there is ample room for further research and development in the L2 education and SLA context. Because the focus of the present dissertation lies on pedagogical implementations rather than pure language acquisition, I will now move on to outlining the conceptualisations of feedback within the domain of (language) teacher education.

2.4.2. Conceptualisations of Feedback in Education

Across different subject areas of the educational context, there are a variety of different theoretical and practical approaches to feedback. Feedback can generally be conceptualised from a behaviourist, cognitivist or socio-constructivist perspective. In the behaviourist understanding, feedback is seen as a tool that reinforces a desirable or inhibits an unwanted behavior (Krause, 2007). Feedback thus traditionally targets a type of behaviour that is to be altered (Hattie & Timperley, 2007; Mory, 2004; Narciss, 2006). From a behaviourist perspective, positive feedback strengthens stimulus-response-associations and thus heightens the likelihood for the desired behaviour to occur. Accordingly, negative feedback (or the lack of positive feedback) is assumed to have the opposite effect, namely to reduce the likelihood

for an unwanted behaviour to occur (Krause, 2007). In contrast, the approach to feedback based on a cognitivist understanding of learning represents the so-called *old* or *traditional feedback paradigm*. Cognitivist learning theories place an emphasis on cognitive processes that occur between a stimulus and response that influence learning. It moves away from focusing merely on visible behaviour and places information processing as well as planning and decision-making processes at its core. Instead of viewing feedback as a reinforcement or inhibition, it is conceptualised as a source of information that learners use to regulate their learning activity and learning process. Feedback definitions by Shute (2008) or Hattie and Timperley (2007) are well-known examples that align with the understanding that feedback equals information transmission. This perspective focuses mainly on the content and the delivery of feedback and dominated most of the feedback literature until recently (Ajjawi & Boud, 2017). In the cognitivist paradigm, feedback is understood as a unidirectional process where the feedback provider (i.e. the teacher) passes on information on the receiver's strengths and weaknesses, thereby focusing on the discrete performance in action (Ajjawi & Boud, 2017). From this viewpoint, the teacher is the active feedback provider and the student is the passive feedback recipient who is dependent on the teacher to assess her or his work. In this *traditional feedback paradigm*, the feedback process is static, monological and information-centered, and constitutes hierarchical roles between the teacher and student (Ajjawi & Boud, 2017; Ajjawi & Regehr, 2019). It also means that the feedback process is often disconnected from tasks and learning activities. This teacher-centered, transmission-oriented model is defined by the type of information within the feedback message itself that is intended to help students improve their learning (Ajjawi & Regehr, 2019).

Newer developments in feedback research have shifted away from a cognitivist towards a socio-constructivist understanding. The latter is termed the *contemporary paradigm* which is rooted in Vygotsky's socio-cultural theories of learning (Vygotsky, 1978; Vygotsky, 1986). The contemporary paradigm approaches feedback from a holistic and process-oriented perspective (Carless, 2015). This socio-constructivist understanding abandons the teacher-centered perspective on feedback. Instead, it places the student as an autonomous and active social agent at its core. Socio-constructivist theories view learning as a self-regulated construction process. Hence, the information a learner perceives from her or his environment is observed, interpreted and processed based on her or his individual pre-existing knowledge, personal beliefs and values, motivational and volitional attitudes and metacognitive abilities (Foerster & Pörksen, 1998; Watzlawick, 1976). Accordingly, feedback is much more than a one-way interaction of

teachers transmitting information to passively absorbent students on respective strengths and weaknesses and outlining ways to improve. Instead, feedback viewed through the constructivist lens places its emphasis on the active role of the student in the feedback process, notably understanding that the student makes sense of the received information from various sources and uses the feedback comments to improve her or his subsequent work or learning strategies (Carless & Boud, 2018). Accordingly, feedback is a multidirectional, cyclical and student-centered process, which is collaborative, interactional and socially constructed (Carless, 2006; Higgins et al., 2002). The future-orientation and inclusiveness of the process turns teachers and students into feedback participants with shared responsibilities (Carless, 2020a). In other words, the present paradigm shift starts to view feedback "as part of an ongoing [educational] relationship between teacher and student" (Ajjawi & Boud, 2017, p. 252). From the perspective of the contemporary feedback paradigm, feedback's main purpose is promote students' self-regulation by facilitating the development of their abilities to monitor, evaluate and regulate their own learning (Nicol, 2010; Price et al., 2010). Consequentially, it is reasonable to assume that feedback can only be effective if the recipients make active use of the feedback information. Carless and Boud (2018) call this student-centered action *student uptake* and argue that this view promotes a repositioning of the feedback concept back in its conceptual roots of cybernetics, systems theory and engineering (Müller & Ditton, 2014b; Narciss, 2006) where it serves to improve work and performance (see chapter 2.4.1). From Carless and Boud's (2018) understanding of *student uptake* within a cyclical feedback process, the recipients develop their interpretations of the received feedback through dialogue, sense making, co-construction and negotiation of meaning in collaboration with the feedback provider. Activities that follow iterative learning processes like two- or multi-stage assignments, for example, are particularly conducive for such a mutual and reciprocal meaning construction (Carless & Boud, 2018). The socio-constructivist understanding, i.e. the contemporary feedback paradigm constitutes the theoretical framework on which I build the upcoming arguments. To understand those better, I will next elaborate on student uptake and what is required for it.

2.4.3. Feedback Literacy

The socio-constructivist reconceptualisation of feedback as a dialogic and relational activity lends supportive evidence that dialogic feedback can be seen as a key strategy for *sustainable assessment* (Ajjawi & Boud, 2018). Indeed, the contemporary feedback paradigm can be positioned within the concepts of *sustainable assessment*, *assessment for learning* (AfL) and

formative assessment. The *assessment for learning* school of thought emphasises the prominent role assessment takes in promoting learning (Broadfoot & Black, 2004; Gipps, 1994; Inbar-Lourie, 2013). It is to be distinguished from *assessment of learning,* which refers to practices that are used to determine achievement levels, for example by means of standardised tests (Inbar-Lourie, 2013). *Assessment for learning* roots in the socio-constructivist theory of learning and considers learners as active participants in a learning community where assessment is used as a tool to empower learners and to improve learning (Inbar-Lourie, 2013). For advancing the learning process, assessment needs to be ongoing, student-related and context-embedded. *Assessment for learning* as a concept recognises the social role of assessment and the power relations within and across assessment procedures, and considers *formative assessment* as a pivotal factor in fostering learning (Black & William, 1998). *Sustainable assessment* has been proposed as a way of conceptualising the purpose of assessment to build lifelong learning capabilities (Ajjawi & Boud, 2018). Boud (2000) defines it as "assessment that meets the needs of the present without compromising the ability of students to meet their own future learning needs" (p. 152). In all three of the above approaches, feedback plays an integral part. To be able to implement sustainable assessment practices, specific knowledge and skills are required. These skills are often subsumed under the generic *assessment literacy* (AL) concept, which refers to the knowledge and skills that are needed for performing assessment-related actions (Stiggins, 1991). A derivative of general *assessment literacy* constitutes *language assessment literacy.* It is contextualised in the field of language teaching and learning and includes both general educational assessment principles and additional components that are specific to language assessment. Although the definitions of *language assessment literacy* (LAL) vary, LAL generally encompasses the unique and specialised knowledge base related to designing and administering language assessments as well as to interpreting, utilising, and reporting language assessment data for different purposes (Inbar-Lourie, 2013). Taylor (2009) proposes a more concrete definition of AL in the language domain by also including aspects of assessment practices that are external to language learning and testing:

> This current trend in thinking seems to be that training for assessment literacy entails an appropriate balance of technical know-how, and understanding of principles, but all firmly contextualised within a sound understanding of the role and function of assessment within education and society. (p. 27)

However, it is not yet conclusive what exactly constitutes basic LAL competence, and who needs to acquire which competences in which context (Inbar-Lourie, 2013). For now, the way

LAL is conceptualised and the depth of expertise it requires seem to indicate that it mainly concerns language testers or policy makers, i.e. high-impact stakeholders. What exactly LAL means for and requires of L2 teachers is not clear and more research in this area is needed (Inbar-Lourie, 2013). That L2 teachers need to develop LAL is thus as of yet an unfounded assumption derived from the concept of general AL. In contrast, the literature and research on AL is more advanced. Overall, general *assessment literacy* can be divided into teacher AL and student AL (Carless et al., 2011). Teacher AL involves the skills needed to help students achieve higher levels of academic achievement, for example through implementing appropriate assessment tools for the respective assessment purpose, scoring and analysing the assessment results fairly and appropriately, using the results for student advancement and providing quality comments and suggestions to improve students' learning processes (Winstone et al., 2017). Student AL includes, among others, the skills they need to understand assessment criteria. Reviewing the constituents of teacher AL and the purpose of *sustainable assessment* makes explicit that feedback skills form an integral part of both. In Boud and Falchikov's (2007) four-step model to develop assessment literacy, for example, developing feedback skills is listed as highly significant. Based on this framework, Price et al. (2010) support this assumption and argue that feedback skills serve the purpose of promoting AL (Price et al., 2010). Thus, feedback can be seen as an essential component of AfL and AL (Carless et al., 2011). Within this AfL context, Carless et al. (2011) include and coin the term *effective sustainable feedback,* which they define as "dialogic processes and activities which can support and inform the student on the current task, whilst also developing the ability to self-regulate performance on future tasks" (p. 397). The following principles maintain that *effective sustainable feedback for learning*

- enhances students' self-evaluative abilities,
- is based on dialogic interaction involving both feedback participants' critique,
- is most powerful in two or multi-stage assessment,
- and makes use of technology to facilitate feedback (Carless et al., 2011).

Accordingly, effective sustainable feedback for learning is co-constructed through (dialogic) and reciprocal negotiation of meaning and requires student uptake (Carless et al., 2011). To allow for sustainable feedback skills to develop, another type of literacy is needed: *feedback literacy. Feedback literacy* (Carless & Boud, 2018; Sutton, 2012) originates in AL (Stiggins, 1991, 1997) and academic literacies (Lea & Street, 1998, 2006) and is based on the notion that

all feedback participants involved in the feedback process share mutual responsibilities (Carless & Boud, 2018; Carless et al., 2011; Carless & Winstone, 2020). It is applicable to all levels of learners, and appears in the context of university students. Just like with AL, *feedback literacy* is comprised of two main constituents that function reciprocally. They are student feedback literacy and teacher feedback literacy, which are both prerequisites for student uptake (Carless et al., 2011). Teacher feedback literacy includes the ability to work with colleagues to establish innovative and student-centered feedback methods (Winstone & Carless, 2019). To be feedback-literate as a teacher also means that she or he is able to design and manage assessment conditions that foster students' feedback literacy development (Carless & Boud, 2018; Carless & Winstone, 2020). Thus, teacher feedback literacy includes the expertise to create conditions that are most conducive for students' engagement with feedback (Chong, 2021; Winstone et al., 2017), and to design feedback practices that nurture student uptake and student feedback literacy (Carless, 2020a, 2020b). Student feedback literacy (Carless & Boud, 2018; Molloy et al., 2020; Sutton, 2012) refers to the skills that students need in order to make good use of feedback processes (Hoo et al., 2021). A feedback-literate student is able to develop capacities in making academic evaluative judgements and actively seeks, generates and uses feedback. Within the academic literacies approach, Sutton (2012) conceptualises feedback literacy as the capability of reading, interpreting and using feedback. According to Sutton (2012), feedback literacy subsumes three dimensions:

> an epistemological dimension, i.e. an engagement of learners in knowing (acquiring academic knowledge); an ontological dimension, i.e. an engagement of the self of the learner (investment of identity in academic work) [; and] a practical dimension, i.e. an engagement of learners in acting (reading, thinking about, and feeding forward feedback). (p. 33)

In a more recent approach to feedback literacy, Carless and Boud (2018) extend Sutton's (2012) concept of feedback literacy by including students' "understandings, capacities and dispositions" to process and use feedback (p. 1315). They identify four interrelated features that underpin student feedback literacy:

- appreciating the value of feedback,
- making judgements in increasingly sophisticated ways,
- managing affective factors productively, and

- taking action in response to feedback (Carless & Boud, 2018).

Thus, Carless and Boud (2018) anchor their perspective on student feedback literacy deeply in the socio-constructivist understanding of learning. The following figure illustrates how these features interact with one another (Carless & Boud, 2018):

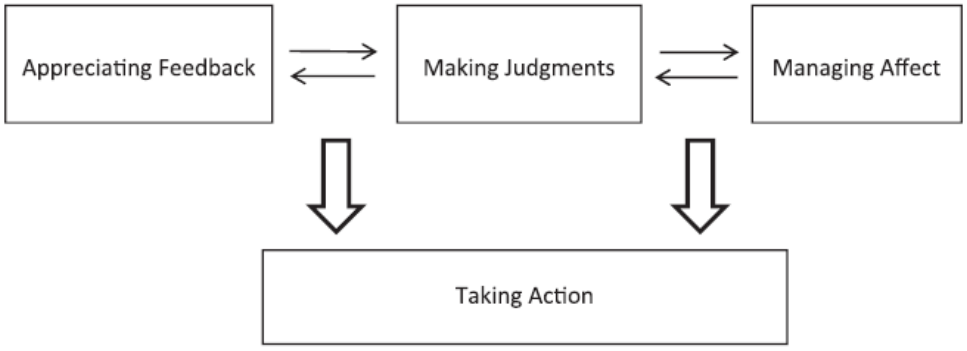


Figure 10 : Features of student feedback literacy

In contrast, behaviours of feedback-illiterate students involve skimming feedback, leaving feedback uncollected (which is partly a consequence of low teacher feedback literacy) (Chong, 2021) and not being able to understand or decipher the language used in the feedback message. Sutton (2012) emphasises this last point by clarifying that "[f]eedback literacy also requires learners and teachers to address the language barriers which inhibit the capacity for learners to understand, interpret and act upon feedback" (p. 49). Indeed, student's language analytical and academic abilities as well as the strategies for using feedback generally influence their engagement with feedback (Burke, 2009). Often, students' feedback literacy is not very highly developed when they enter university. Although university students have been exposed to feedback for their entire educational career, there is evidence that the frequency at which they received feedback in school may have been relatively low (Bond et al., 2000). In addition, it is likely that the little feedback that they had received related to the students' self (also called *personal level*, c.f. Hattie & Timperley, 2007) or remained on the level of corrective feedback (Blöte, 1995). These types of feedback, especially when provided in isolation, tend to have the least impact on students' learning (Hattie & Timperley, 2007). Related to this is the fact that research continuously shows that educators' perception of the quality and quantity of their feedback is more favourable than their learners' perception thereof (Carless, 2006; Price et al., 2011; Urquhart et al., 2014). Indeed, there is evidence that suggests that the feedback models students were exposed to during their schooling career may not have been the most conducive to developing their own feedback literacy. Studies on student perceptions of feedback including

how students receive and respond to feedback such as those conducted by Weaver (2006) and Burke (2009) show that, in many cases, students enter higher education with little knowledge on how to act on feedback. Weaver (2006) found that over 50% of the university students in his sample had not received guidance on how to understand and use feedback at university itself. In addition, his findings indicated that three-quarters of students had not received any guidance on using feedback before entering university. In the study by Burke (2009) to identify what kind of guidance for using feedback Humanities students bring to university, she found – in contrast to Weaver – that almost 40% had received guidance for using feedback before entering higher education. However, students commonly confused *actual feedback* with *guidance on how to use* feedback. Burke's 2009 research supports Weaver's 2006 findings that the majority of students entering higher education lack student feedback literacy because they do not possess *strategies* to act on feedback. However, the acquisition of such *strategies*, i.e. student feedback literacy, can be promoted through carefully planned scaffolding, for which multi-stage assignments (e.g., portfolios) are particularly conducive (Carless & Boud, 2018; Mutch et al., 2018; Sutton, 2012). Student feedback literacy can be trained through

- collaborative interaction (feedback dialogue) within socio-constructivist principles,
- peer feedback (exposure to work of peers), which involves a collaborative learning process (Tai et al., 2018). Through peer feedback, a learner's knowledge about quality (*evaluative judgement*) becomes linked through repeated exposure to work, the consideration of standards (implicit or explicit) and the need to justify decisions which help strengthen judgements in relation to particular disciplinary genres within a community of practice,
- understanding and applying rubrics (which represents a move towards the enactment paradigm where students are seen as active agents),
- exposure to exemplars (ibid.).

Mostly analogous recommendations are made by Hoo et al. (2021) based on the findings of their recent study on student feedback literacy. They propose three phases of learning to promote student feedback literacy: 1) self-awareness via self-assessment and feedback; 2) inquiry and negotiation of multi-source feedback (e.g., self and peer); and 3) putting plans into action and monitoring progress in relation to original plans (p. 11). In summary, if students are feedback-literate, uptake of feedback is more likely to occur. However, regardless of the sophistication of students' feedback literacy, for learners to take up their teachers' feedback it

is essential that they have the communicative skills to participate in a co-constructive dialogue and understand the teachers. Likewise, the way a feedback message is constructed and framed is a deciding factor for students to understand it. Some of the above considerations imply that teachers' communicative skills or overall language use may not only promote to facilitate feedback, but may even be partly responsible for its success. Hence, one can argue that language plays an important role in the entire feedback process, and that successful feedback demands the activation of specific (or specialised) communicative strategies and language competence by all feedback participants involved.

The conceptualisation of student feedback literacy and ways to develop it (Carless & Boud, 2018; Sutton, 2012) is still in its infancy (Chong, 2021). Thus, the above pedagogical and didactic recommendations for fostering feedback literacy development are yet underinvestigated and need further empirical evidence. Nevertheless, the feedback literacy concept provides a suitable theoretical framework for the present dissertation. In sum, the importance of L2 teachers' oral profession-related feedback skills becomes apparent in socio-constructivist approaches and the understanding of the central role of meaningful interaction and the negotiation of meaning in the L2 learning process. This school of thought focuses on exposing learners to large amounts of spoken language and encouraging them to engage in spoken interactions (Hughes, 2011). Such spoken interactions may include feedback conversations that align with the socio-constructivist approach to feedback, and student and teacher feedback literacy. This dominance of spoken interaction in the classroom and its role in the feedback process, paired with the need for designing appropriate assessment instruments to monitor and evaluate oral teacher language competence, results in the need to gain a better understanding of how oral teacher language competence can be fostered and assessed. I will now proceed to discuss the latter from a theoretical perspective before I turn to focus on the challenges of assessing L2 oral language competence and/or performance and its implications for language assessment in general.

2.5. Assessing Oral Language Competence

As outlined in the previous section, the concept of competence is complex. This poses considerable challenges when it comes to assessing competence in its entirety. If one seeks to assess *oral language* competences, the complexity of the nature of speaking per se intensifies these difficulties. Indeed, speaking is said to be the most difficult language skills to assess

reliably because there are a multitude of factors that can influence an assessor's impression of how well a learner can speak a language (Luoma, 2009). In combination with the expectation that assessments should produce fair, valid, reliable and objective scores that are appropriate to the context and purpose of any given test, assessing oral L2 competence presents an endeavour with many requirements and challenges (Luoma, 2009). This chapter is dedicated to discussing these challenges, both from a more general language-testing perspective as well as from a more contextualised, "local" point of view with reference to L2 teacher education and teacher language competence. The first section provides an overview of the necessary fundamentals of language testing. That there are many ways of assessing (oral) language is described in the following sections. These include the presentation of a few relevant assessment and scoring types, and an excursus on implications of the specific nature of speaking on oral L2 assessment. Subsequently, different ways of how L2 teachers' (oral) language competence can be (and currently are) assessed are outlined and discussed. Finally, considerations on scoring oral L2 competence and/or performance are presented, including challenges related to the reliability of human raters. Ways of how these challenges can be mitigated and a brief synopsis of setting standards as steps that follow the rating process conclude this subchapter.

2.5.1. Fundamentals of Language Testing

Language tests are instruments that elicit an L2 learner's language performances. These performances serve as grounds to identify a learner's language ability by drawing inferences on the underlying competence as conceptualised in the relevant competence model and operationalised in the test purpose (Douglas, 2010). A multitude of language test types exist that are designed to evaluate a language learner's *communicative ability* (cf. Harsch, 2016 for an overview), however, measuring something as complex and intangible as the knowledge of language (Douglas, 2010) is challenging. As Spolsky (1995) poignantly describes:

> The fundamental flaw of objective modern language testing has been to presuppose that language proficiency is measurable and unidimensional […]. Language proficiency is more like pain, the external assessment of which has many analogous properties: it varies from person to person, from context to context, and can only be inferred from self-report and the observation of impaired performance. […] To assume that all this complexity can usefully and meaningfully be squeezed into a single number or a single point on a unidimensional scale, is, on the face of it, absurd. […] Only the most elaborate test

> batteries, with multiple administrations of multiple methods of testing the multiple traits or abilities that make up language proficiency, are capable of producing rich and accurate enough profiles to be used for making critical or fateful decisions about individuals. (p. 357-358)

Essentially then, language tests aim to assess language competence and/or proficiency based on the inherent discrepancy of competence and performance. Because competence can only be inferred based on observed performance and therefore relies on interpretations (see chapters 2.1 and 2.2.1), a certain level of uncertainty when assessing language competence always remains. The following sections outline inherent dilemmas of language testing and describe practical implications based on pragmatic approaches to the fundamental flaws as outlined by Spolsky (1995). I thereby build on the view that *language ability* constitutes the expression of language use, i.e. of reading, writing, speaking and listening (integrated), and that *language ability* is therefore the object of measurement (Bachman & Palmer, 1996; Douglas, 2010; Fulcher & Davidson, 2007; Lado, 1961).

2.5.1.1. Quality Criteria of Language Tests

To ensure that a language test produces fair and valid results, it must meet rigid quality standards. There are different ways of how these standards can be approached. In classical test theory, the three primary quality criteria *reliability*, *validity* and *objectivity*, and the four secondary quality criteria *standardisation*, *comparability*, *ecology* and *usefulness* are understood to assess the quality of a test (Lienert & Raatz, 1998). Additional quality criteria, especially for communicative language tests, constitute *distinctiveness*, *practicality* (also referred to as *feasibility,* a criterion that is particularly relevant to performance assessment, cf. Council of Europe, 2001), *interactiveness*, *authenticity*, *transparency* and *washback* effect[11] (Harsch, 2016). Due to their particular relevance to the present research, I will briefly outline the three primary quality criteria. Reliability as the first primary quality criterion measures the

[11] Washback (Alderson & Wall, 1993) or backwash (Hughes, 1993) refers to the effect tests have on teaching and learning. Such effects can be either positive (Taylor, 2005) or negative (Brown, 2004). Washback concerns aspects of consequential validity and is often used when referring to the impact a test has or may have on the precursory course and/or teaching leading up to the test (McNamara, 1996). Positive washback effects may occur when a particular *teaching to the test* approach – e.g., preparing learners for communicative language tests – indirectly prepares learners for real-world communicative tasks beyond the test itself (Taylor, 2005). Negative washback refers to possible harmful impacts on teaching programmes or curricula, e.g., through constraining the course content to mere test preparation rather than preparation for real-world communicative tasks (Brown, 2004). Another way washback could be negative is if a classification error occurs, for example when classroom instruction prepares learners for a pen and paper test that seeks to assess communicative competence. Such teaching to the test may have counterproductive effects on real-world language acquisition and use (Bachman & Palmer, 1996).

accuracy to which the test responses or results can be reproduced should the test be retaken (Grum, 2012). The more accurately a test can capture the characteristics it seeks to measure, the higher its reproducibility. In other words, a reliable test – especially in psychometric testing – is one in which the test performance remains constant both across different learners of the same ability(ies) as well as different administrations of the same test with the same testers (Levi, 2012). Test reliability measures include, among others, *interrater* and *intrarater reliability,* which can be identified through statistical calculations of reliability coefficients (see chapters 2.5.4.4 and 5.1.1). Validity as the second primary quality criterion measures the appropriateness of the inferences made based on the test performance (Douglas, 2010). It indicates whether a given test indeed measures what it intends to, i.e. whether it measures the competences and skills that build the test construct. Validity is considered the most important – and at the same time most notoriously difficult – primary quality criterion and sets the minimum standard for a conclusive test. There are different types of validity (e.g., construct validity, face validity, content validity, etc.; cf. Grum, 2012 or McNamara, 1996 for a detailed overview). To measure test validity, one needs to be aware of the difference between test validity and test validation (Xi & Sawaki, 2017). While validity refers to the "theoretical notion that defines the scope and nature of validation work, […] validation [constitutes] the process of developing and evaluating evidence for a proposed score interpretation and use" (ibid. p. 194). It is the conceptualisation of validity that determines the type of evidence to collect and how and to what extent to execute the test validation in context (Xi & Sawaki, 2017). In contrast to measuring the different types of validity separately or to prioritising one type over another, an argument-based approach proposes to address validity more holistically. Essentially, an argument-based approach includes collecting and coherently analysing various types of evidence for and against a proposed score interpretation to test the assumptions that "test scores and other related information provided to users are *relevant*, *useful*, and *sufficient* for making intended decisions; the decision-making processes are appropriate; and the assessment process does not incur any negative consequences" (Xi & Sawaki, 2017, p. 198, italics in original). This holistic approach originates in Kane's framework to test validation in education measurement (Kane, 1992; Kane et al., 1999) and its corresponding spin-offs in language testing for example by Bachman (2005), Bachmann and Palmer (2010), or Chapelle et al. (2008). Objectivity as the third primary quality criterion stands for the test's generalisability and the respective independence between the evaluation and the evaluator or the test instrument. If different testers and raters reach the same judgements about the same test takers' test performance, a test can be identified as

objective (Grum, 2012; Lienert & Raatz, 1998). Aspects that add to the explanatory power of this quality criterion are the objectivity of the test administrators and test evaluators, and the independence of the test result interpretation achieved through systematic and consistent allocation of test scores to competence levels, grades and the like (Grum, 2012; Lienert & Raatz, 1998). Consolidating the above, Bachman and Palmer (1996) suggest that the criterion *usefulness* is essentially a compound of the test reliability, construct validity, authenticity, interactiveness, impact and practicality. Indeed, *usefulness* constitutes an all-encompassing umbrella-criterion that ultimately decides on a test's systemic importance:

> The most important consideration in designing and developing a language test is the use for which it is intended, so that the most important quality of a test is its usefulness. […] although there is a tension among the different test qualities, this need not lead to the total abandonment of any. It is our view that rather than emphasising the tension among the different qualities, test developers need to recognise their complementarity. (Bachman & Palmer, 1996, p. 17)

Bachman and Palmer (1996) thereby underline that, while there seems an implicit hierarchy between the primary and secondary quality criteria, they are also very much interrelated. For example, as much as objectivity is a prerequisite for reliability, reliability is also a prerequisite for validity. Most importantly, though, one needs to keep in mind that even if a language test fulfils all the (complementary) test quality criteria, tests merely allow for making inferences about a learner's *language ability* based on a given observed performance. There always remains a certain extent of uncertainty about what truly constitutes a learner's L2 ability. Thus, test results always need to be treated and interpreted with caution, especially when high stakes decisions are made based on the results (Douglas, 2010).

2.5.1.2. Authenticity

Language never happens in a vacuum but is always used for a specific purpose and related to a specific context and situation (Douglas, 2010; Lado, 1961). Accordingly, L2 ability can and should not be separated from the assessment context, because assessments are just as much locally-situated (i.e., with embedded test tasks as operationalised in the scale) as a real-world L2 use (Chalhoub-Deville, 2003). It is therefore central that language tests do not only draw on isolated vocabulary or grammar knowledge removed from the relevant context. Indeed, there is consensus in the literature that speaking assessment should employ a scenario-based approach that allows the relevant context to be represented in the test tasks (Seong, 2017). As Douglas

(2010) points out: "[…] if the test purpose is to make inferences about a learner's language ability in some communicative context, then the test should provide relevant contextual information" (p. 21). Relevant contextual features applicable to all situations of language use encompass for instance setting, participants, or genre (Douglas, 2010; Hymes, 1974). In order to ensure that a test taker interprets the contextual cues of a test task as intended, and consequently produces the language production the test is designed to elicit, the contextualisation of each test item needs to be very specific. The aim is to reduce the scope for interpretation to a minimum so that the test takers behave in a way that allows for a reliable test result interpretation. The high relevance of context setting in language testing is closely connected to the high relevance of authenticity (Douglas, 2010). Indeed, authenticity stands in direct relation to the validity and usefulness of a test and references the relationship between the specifics of a test and real-world language use (i.e. target language use (TLU)). Bachman and Palmer (1996) define authenticity as

> the degree of correspondence of the characteristics of a given language test task to the features of a TLU task. […] Authenticity thus provides a means for investigating the extent to which score interpretations generalise beyond performance on the test to language use in the TLU domain […]. This links authenticity to construct validity. (p. 24)

Elder and McNamara (2016) postulate that a "key requirement for authenticity in [language for specific purpose] testing […] is establishing what communication entails in the particular context of concern" (p. 148). Authenticity in language testing is comprised of *situational authenticity* (the degree to which features of real-life tasks are reproduced in test tasks) and *interactional authenticity* (the extent to which test takers engage cognitive processes they would employ in real-life language use when completing a test task) (Kitney & Morgan, 2019). A test of high situational authenticity requires the test takers to "respond to contexts which simulate 'real life' in terms of criterial parameters without necessarily replicating it exactly" (Weir et al., 2013, p. 212). If a test is of high interactional authenticity, the test takers are required to engage in the cognitive processes that "are representative of, and offer adequate coverage of, the cognitive processes which would prevail in a natural (i.e. non-test) context" (ibid. p. 97). Situational authenticity needs to be distinguished from *genuineness*. In language testing, *genuineness* "is a property of a spoken or written text and results from the text having been produced in an actual communicative situation" (Douglas, 2010, p. 25). An example of a *genuine* artefact in a language for specific purposes (LSP) test for assessing teacher language competence would for instance be a piece of writing that a student from the target level

produced in the TLU domain, i.e. in an actual L2 lesson. Even though the *genuineness* of this artefact would contribute to the LSP test's authenticity, it would however not be able to guarantee authenticity in itself. To increase the level of authenticity, Hymes' contextual features would need to correspond to the real-world equivalent, too (Hymes, 1972). Another way of enhancing authenticity in test tasks is by following the Ladoesque (1961) notion of assessing language ability by means of grouping together the integrated skills[12] speaking, listening, and writing because they naturally occur in tandem in the real-world tasks that test-takers are likely to undertake (Plakans, 2013). Thus, integrating stimuli taken from real-world contexts into test tasks to elicit language productions can contribute to enhancing the level of authenticity of a test. Vignettes are one way of doing so. A vignette is an artefact that typically simulates a complex real-world scenario by means of a short story about hypothetical characters in specific circumstances (Finch, 1987). It can take many different forms, ranging from textual descriptions (text-vignettes) to audio-visual representations of a situation (video-vignettes), and from short written prompts to live events (Hughes & Huby, 2002). While in contrast to genuine artefacts vignettes contain fewer complexities than real life, they enable the depiction of a holistic picture of the scenario at hand – especially in the case of video-vignettes. Indeed, "viewing videos of behaviors and interactions [are rhetorically powerful contributors] for understanding nuances of social relationships, kinesics, proxemics, prosodics and other situated parameters of human interactions" (Goldman et al., 2007, p. xi). Such information-enriched video-scenarios can thus be used to increase the level of authenticity and at the same time decrease the room for interpretation. By diminishing the interpretation slack, comparability between test tasks and test takers' responses can be increased. This prerequisite ideally leads to the test takers reacting as intended, thus allowing for more comparable test results. At the same time, vignettes set up "a situation in which there is no one "right" answer, and [they are] flexible enough that individuals from different groups […] can identify with the story and bring their perspective forward in discussions of solutions" (Campbell, 1996, p. 2).

Ensuring a high level of authenticity in language tests is a noble quest. Challenges include ensuring the reliability of the test scores and mitigating the field of tension of offering a broad

[12] According to Hallet and Königs (2010), the concept of *integrated skills* refers to the linkage of two or more linguistic skills in the context of language teaching and learning. The term 'Integration' (lat. Integratio = renewing, restoring, the act or process of making whole or entire) points to the fusion of the linguistic skills to a "whole" (verbal) communication. Similarly, Oxford et al. (1994) define language skill integration as follows: "It involves linking the four language skills of listening, reading, speaking and writing with the intent of emphasising real, meaningful communication. It also involves integrating supportive skills such as grammar, pronunciation, and vocabulary development, as well as the general area of culture, which is inextricable from language" (p. 257).

enough range of authentic real-life tasks to elicit skills relevant to the test taker while maintaining comparability of test-task interpretation and test-task responses (Kitney & Morgan, 2019). Indeed, despite the efforts that can be undertaken to increase a test's level of authenticity, it is also debated whether tests can ever be completely authentic to real life at all. Considering that we are socialised to know what to expect of a test and that tests themselves are not part of the "real world", complete authenticity in a test may seem like an illusion (Pill, 2019). Often, the quest for increased authenticity is also limited by mere pragmatic constraints of language testing, or for reasons of practicality[13] of the test administration. It is instead argued that tests have their own authenticity, and that there is a notable difference between *authentic tests of language* and *tests of authentic language* (Lewkowicz, 2000). Nevertheless, authenticity in language testing is highly relevant, not only because of the above reasons but also, as Bachman and Palmer (1996) state,

> because of its potential effect on test takers' perceptions of the test and, hence, on their performance. […] It is this relevance, as perceived by the test taker, that we believe helps promote a positive affective response to the test task and can thus help test takers perform at their best. (p. 24)

Douglas (2010) rightly adds that authenticity in language testing can only be taken so far. However, it is nevertheless necessary that test developers make the "effort to provide a context for language use in our tests to help ensure that the interpretations we make of the test takers' performances will be valid" (p. 26). Pragmatic approaches such as integrating video-vignettes as authentic (or even genuine) material can contribute to such efforts. The extent to which these aspirations must be and can be pursued not only depends on the test construct, but also largely on the test purpose and type of language test at hand. In the following section, I outline types of language tests that are relevant and applicable for assessing speaking and, more precisely, for testing oral teacher language competence.

2.5.2. Communicative Language Testing

Language tests appear in different forms. While the test construct and test purpose generally determine their type (i.e., proficiency, achievement, placement, diagnostic, or aptitude tests, cf. Douglas, 2010; Harsch, 2016), they can also be categorised according to a variety of criteria

[13] Bachman and Palmer (1996) define practicality as the difference between the required and available resources for the test development and use.

(i.e., they can be formative or summative, norm-oriented or criteria-oriented, high stakes or low stakes, large-scale or small-scale, etc., cf. Harsch, 2016). For communicative language tests – a type of test that is particularly significant when it comes to assessing oral L2 competence – a classification according to the following three categories is especially meaningful (ibid., see also chapter 2.5.2.2):

- indirect tests (e.g., competence tests that elicit relevant language productions in order to draw inferences regarding the underlying competence),
- semi-direct tests (where the language ability is generally tested on a global scale and by means of integrated test tasks, however the test tasks are not directly placed in an authentic setting. Instead, authenticity is simulated), and
- direct tests (e.g., performance tests where the test task completion requires exactly the language ability that is sought to be tested; hence, the ability itself is the object of the test).

Communicative language tests ground their rationale in the theory of communicative competence that arose from the communicative turn in the 1970s (McNamara, 1996). Indeed, Hymes' theories on communicative competence (1972) have greatly influenced (performance-based) communicative language tests (McNamara, 1996). Communicative tests assess a learner's

> ability to use language for communication in specific contexts, involving productive language either through meaningful input for the test taker to comprehend or interpret, or as meaningful output generated by the test taker. (Douglas, 2010, p. 69)

The communicative paradigm is the underlying rationale for this approach. It considers that a learner not only needs language knowledge but primarily also *communicative competence* or the ability for language use (Hymes, 1972) to become a competent language user. Hence, communicative language tests are designed to elicit language performance in relevant contexts of use rather than to test knowledge of isolated aspects of language such as grammatical structures or phonology (Douglas, 2010). An example of test tasks typical of the latter would be *discrete-point test tasks*, or, in speaking assessment in particular, *structured speaking tasks* (Luoma, 2009). These type of tasks generally focus on isolated points of grammar, vocabulary, syntax etc. (i.e. structuralist approach to language testing, ibid.). In contrast, *integrative test tasks* or *open-ended speaking tasks* align with the communicative approach and require test takers to process a number of language aspects simultaneously in order to respond (Luoma,

2009). *Open-ended speaking tasks* allow room for different ways of completing the task and typically call for extended responses (Luoma, 2009). They may vary according to the discourse type they aim to elicit (e.g., description, instruction, explanation, role-play, etc.) and according to the amount of test takers and interlocutors involved (Luoma, 2009). There are many more sub-types of test tasks of which Sari Luoma's book *Assessing Speaking* (2009) provides a comprehensive overview. In communicative language testing, open-ended test tasks are more common and appropriate. As previously mentioned, it is important to remember that underlying the observed language performances in a language test lie the components of communicative language ability (Douglas, 2010). This means that even in a direct test that seeks to measure a test taker's speaking ability, what the test actually measures is *language ability* revealed through the spoken medium (Douglas, 2010). I will now proceed to outline ways of language testing that lend themselves to measuring the construct of interest at hand: oral L2 teacher language competence.

2.5.2.1. Language Assessment for Specific and Professional Purposes

The differentiation between *general language tests* and *specific purpose language* tests grounds on a long-lasting history in language teaching and assessment (Douglas, 2010). Generally, the distinction is based on the idea that the purposes for learning are not distinctively specified in *general language tests*, whereas in *specific purpose language* tests (LSP) they are. Specific purpose language teaching and assessment is considered a distinct branch of applied linguistics. However, some researchers criticise the theoretical distinction between general and specific purpose language teaching and assessment as no longer being viable (Douglas, 2010). One of the main reasons for this rejection is that no language test or course can be designed without a purpose (ibid.). Even though the contextual purposes of language teaching and language assessment are situated on a continuum where the degree of specificity may vary, Douglas (2010) argues that the distinction between *general* and *specific* purpose has become blurred. Knoch and Macqueen (2020) propose a new approach to LSP testing by broadening the concept of LSP to *language assessments for professional purposes (LAPPs),* thereby including assessments for professionals (*domain insiders*) and assessments for laypeople (*domain outsiders*). In their 2020 publication, they define LAPPs as language tests that are conducted "in relation to participation in professional settings" (Knoch & Macqueen, 2020). Thus, LAPPs are

> [a]ny assessment process, carried out by and for invested parties, which is used to determine a person's ability to understand and/or use the language of a professionally-oriented domain to a specified or necessary level. (ibid. p. 20)

Instead of replacing LSP tests with LAPP tests, Knoch and Macqueen identify the differences, similarities and overlaps between the two. While LAPP solely refers to language testing in the professional domain, i.e. the *Target Language Use (TLU) domain* (e.g., the medical and health industry, or the aviation industry), LSP also includes language assessments for academic purposes (often in relation to the target language English, abbreviated as EAP) (Knoch & Macqueen, 2020). Indeed, the TLU domain is a core characteristic of LAPP tests. For instance, LAPPs are often used to identify so-called *proficiency thresholds,* a cut-off point that determines whether someone is considered to meet the minimum standard of required language ability in order to cope with the language demands of that particular professional domain (Knoch & Macqueen, 2020). Of fundamental concern to LAPPs are whether the relationship is established between the "evidence" collected and the TLU domain to which these scores are claimed to be relevant (Douglas, 2000), and whether this evidence provides a sufficiently reliable basis to make inferences about a test taker's language ability in the TLU domain (Knoch & Macqueen, 2020). This relationship can be characterised in terms of its authenticity (how close is the evidence to the language of the target domain?), its specificity (to what extent is the evidence focused on the domain, or to what degree can it be considered domain-specific?), and its validity (how meaningful and fair are the scores if they are used in relation to the target domain?) (Knoch & Macqueen, 2020). LAPPs face the challenge that they need to account for two main aspects that determine a person's successful participation in a professional domain: one, the participant's proficiency in the target language(s) and two, the participant's professional knowledge for a particular occupational role (Douglas, 2000). Indeed, the issue of central concern is "the degree of separability of language ability and professional knowledge in assessments that explicitly seek to categorise people according to language ability" (Knoch & Macqueen, 2020). LAPPs thus often assess test takers in non-language abilities, but since such tests are mostly delivered in a standard language like the L2, the separation between the non-language construct from the language-based method of assessment becomes almost impossible (Knoch & Macqueen, 2020). Hoekje (2016) highlights this problem by pointing out that the legal requirements of high-stakes LSP tests often "demand more consistency in separating 'language' from professional practice" (p. 292). This creates a dissonance between the theories of communicative competence that underlie LSP tests and the contradictory demand of

separating content from language. This then stands in opposition to the concept of communicative competence. Central issues for LSP and LAPP tests are thus bringing in accordance the assessment of

> language on a communicative task, that is, an adequate theorization of the role of content and meaning, the domain (context) of use, the role of the interlocutor in interaction, and the contribution of the nonverbal. (ibid. p. 292)

In addition, and related to these issues, is the fact that LSP or LAPP performance tests are almost inevitably confronted with coexisting "double constructs". The Occupational English Test (OET), for instance, assesses both healthcare communication and the LSP construct of communication for healthcare. Overlapping constructs such as these make an LSP or LAPP test particularly vulnerable to validity challenges such as construct underrepresentation and construct irrelevant variance (Hoekje, 2016; Messick, 1994). To better understand these problems, it is helpful to consult the categorisation that generally applies to LAPPs. Indeed, the two types of LAPPs are 1) stand-alone LSP assessments and 2) assessment conducted in LSP courses (Knoch & Macqueen, 2020). The primary purpose of stand-alone assessments is to predict behaviour in the *real-world* workplace. Any LAPP stand-alone assessment may fall somewhere on a spectrum ranging from being mainly language-focused to being mainly content-focused. Additionally, stand-alone assessments are often administered in form of large-scale, standardised tests that fulfil a gate-keeping role (e.g., allowing (or disallowing) access or registration to a profession, granting visa etc.). Course assessments are similar to stand-alone LSP assessment with reference to the language-content spectrum (Knoch & Macqueen, 2020). While LSP/LAPP stand-alone assessments provide a predictive view on potential performance in the target domain, course assessments provide a retrospective view by measuring achievement that results from a prior period of learning and by drawing implications for future learning or performance (Knoch & Macqueen, 2020). Both LAPP types highlight that "there are […] many 'professional knowledge' assessments which implicate language and 'language' assessments which implicate professional knowledge" (Knoch & Macqueen, 2020, p. 45). This variety indicates the difficulty that exists when it comes to measuring language that is of a specific kind or purpose, resulting in challenges such as

> [t]he fact that tests of general language proficiency are routinely used to classify the workplace 'readiness' of people's language abilities, [which] shows that, in practice, there is a pervasive view that professional language ability can be adequately demonstrated through general language use. (ibid., p. 45)

However, a test score alone can by definition not represent workplace readiness, or "the knowledge base provided by extensive acculturation into a specific workplace culture and the opportunities there for the acquisition of a Secondary Discourse" (Hoekje, 2016, p. 296). It is thus pertinent that LSP test scores are used and interpreted according to their relative explanatory power. This means commonly accepting and promoting the idea that such test scores much more likely indicate that a test taker has reached a threshold readiness to begin the workplace acculturation process. This is also to avoid promoting the perception that the completion of the acculturation process can be reflected, let alone measured, by a passing score (Hoekje, 2016).

2.5.2.2. Performance Tests

Performance tests – also referred to as task-based assessment (Douglas, 2010) – present a possible test format for LSP or LAPP tests. They are a form of communicative language test involving complex test tasks with clear communicative goals that reach beyond the mere purpose of displaying a language skill (Douglas, 2010). The main purpose of performance tests is to measure communicative language ability under *performance conditions* in (near)authentic settings that reproduce the complexity of language use outside of the classroom context (McNamara, 1996). Thus, performance tests are characterised in terms of the extent to which they simulate the particular real-life context (McNamara, 1996). The focus of such tests lies more on the process of a performance "and what the performance reveals about the *underlying state of language knowledge*" (ibid. p. 6, italics in original) rather than a performance end goal. With reference to performance tests in occupational settings and building on Slater's earlier work (1980), Jones (1985) discusses three formats in which performance assessments can be realised: *direct assessment, work sample methods* and *simulation techniques* (see also chapter 2.5.2). *Direct assessment* involves direct performance observations in the authentic occupational setting (e.g., the real-life classroom in the case of L2 teachers). Since language behaviour is very complex, observations over an extended period would be necessary in order to gain sufficient grounds on which to make valid judgements about a candidate's language ability. This fact alone makes *direct assessment* very time-consuming. A more practicable alternative is the *work sample method*, which also involves observations in the real-life context but sets standardised and controlled test tasks (Jones, 1985; McNamara, 1996). Finally, *simulation techniques* involve test tasks that simulate a situation and employ those simulations in tests outside of the actual real-life context. Thus, the test tasks and test setting are more abstract, and the "performance on the task is […] used to predict performance on similar real-

world tasks" (McNamara, 1996, p. 14). An example of a *simulation task* would be a microteaching sequence. Microteachings usually involve filming short, targeted teaching sequences that include a focus on one or more specific aspects of teaching (e.g., providing feedback, giving instructions etc.) (Kennedy & Lees, 2016). While the forms of performance assessment may differ in terms of their practicability and authenticity, they have in common that the elicited performances are usually scored by means of judgements against a rating scale. Ideally (but not necessarily), the rating scale corresponds to real-life standards of the criterion in question. This poses challenges, such as determining the relationship between linguistic and non-linguistic factors in performance that is to be reflected in the corresponding rating scale (McNamara, 1996). Jones (1985) emphasises this by explaining that *language* in itself is always only one of several factors that language performance tests assess:

> the overall criterion is the successful completion of a task in which the use of language is essential. A performance test is more than a basic proficiency test of communicative competence in that it is related to some kind of performance task. It is entirely possible for some examinees to compensate for low language proficiency by astuteness in other areas. (p. 20)

The question is, then, what other factors need to be considered and included in rating scales, whether it is equitable and ethical to assess these factors, whether these factors are operationalisable in the first place, *how* these factors should be assessed, and to what extent the inclusion of such factors impacts on the feasibility of the overall test (McNamara, 1996). In this context, McNamara (1996) distinguishes between *strong* and *weak* L2 performance tests (cf. also Knoch & Macqueen, 2020). Based on previous work by Messick (1994), McNamara (1996) suggests that both types represent a pole on either end of a continuum. Each pole reflects the extent to which assessment criteria of performance tests mirror indigenous real-world criteria or formal linguistic criteria of language performance (McNamara, 1996). In *strong* performance tests, the test items denote real-world tasks and the elicited performance is primarily

> judged on real-world criteria, that is, on the fulfilment of the task set […]. Such a test thus involves a second language as the *medium* of the performance; performance of the *task* itself is the *target* of the assessment. (Messick, 1994, p. 43)

It is especially in *strong* performance tests where non-linguistic factors play a central role (Jones, 1985). This means that adequate L2 proficiency is not a sufficient prerequisite for

successfully completing a test task. Instead, other non-linguistic factors are essential predictors. In Hymes' (1972) terms, *strong* performance tests thus fully and explicitly integrate language knowledge and ability for use. The closest forms of *strong* performance tests would be either *direct assessments* or *work sample methods* as outlined above (McNamara, 1996). In contrast, *weak* performance tests focus on *language performance* by means of implementing real-world tasks in a test setting removed from the real-world context. Most performance assessments are of this type; they are however somewhat misleading because they superficially imply a *strong* performance assessment:

> [t]he candidate is required to perform on a task which may represent tasks he or she may subsequently face in the real world; however, the capacity to perform the *task* is not actually the focus of the assessment. Rather, the purpose of the assessment is to elicit a language sample so that second language proficiency, and perhaps additionally qualities of the execution of the performance, may be assessed. (McNamara, 1996, p. 44)

The Occupational English Test (OET) is a prominent example of a *weak* LSP performance test as introduced above (see 2.5.2.1). While *weak* performance tests employing approaches such as *simulation techniques* are the most economical and feasible to implement (Jones, 1985), they also give rise to challenges such as questions of validity and the relationship between the test, the test performance and the criterion. Indeed, validity issues related to generalisability across different relevant task types – meaning the issue of the representativeness of the sample performances elicited in a test – and validity issues concerning replicability concern all forms of performance assessments (McNamara, 1996). As Linn, Baker and Dunbar (1991) rightly caution:

> Simply because the [direct,] performance-based measures are derived from actual performance or relatively high-fidelity simulations of performance, it is too often assumed that they are more valid than multiple-choice tests. (p. 16)

Knoch and Macqueen (2020) reflect on this issue within LAPP contexts (see chapter 2.5.2.1). On the basis of the communicative paradigm in language testing Bachman (2002a, 2002b) strongly argues for the consideration of context in test performance. However, Chalhoub-Deville (2003) argues that researchers see performance in context as different from the abilities underlying performance when referring to competence. In other words, the performances elicited through the performance assessment are test-situation-performances in their own right, but they should be transferrable to the performance required in real-life contexts. She postulates

"individual ability and contextual facets interact in ways that change them both" (p. 369). This view supports Swain's (2001) assumption that language abilities are a characteristic of an individual, and that these characteristics are transferable to other contexts. Accordingly, only if this distinction is maintained can (psychometric) generalisations about language abilities be made to other contexts. This assumption calls for the view that the purpose of performance tests should be to interpret the test performance and predict future performance in other contexts rather than to measure language proficiency at a discrete point in time (Levi, 2012). Finally, valid interpretations of test performance inferences on the underlying competence, however, require appropriate and various types of evidence (Douglas, 2010). Such evidence includes, among others, numerical empirical evidence collected both *a priori* and *a posteriori* (McNamara, 1996). The degree to which generalisations are possible based on the evidence, however, partly depends on the assessment context, on what performances are considered to be authentic, and on whether a broad or narrow definition of performance tests is adopted (McNamara, 1996). As this introduction shows, performance assessment is a particularly complex type of language test. Its rich setting and complex test construct renders performance tests particularly vulnerable to vast variability (McNamara, 1996). Due to this complexity, it is also of central importance to make careful considerations with reference to a performance test's practicability and *feasibility* (Council of Europe, 2001). I will now proceed to outline the ways in which the specifics of oral language competence and oral language performance affect considerations and practices within language (performance and/or competence) assessment.

2.5.2.3. Specifics of Oral Language Assessment

The complex nature of performance assessment (see chapter 2.5.2.2) and oral language production render speaking performance assessment to be a particular challenge for language testing (Luoma, 2009). A number of factors need to be considered when designing oral L2 assessments, be they competence or performance assessments. First, as mentioned above (see chapter 2.2.4), language tests need to build on a clearly defined test construct, which should root – depending on the type of language assessment – in an explicit theory of language ability or theory of language performance (Douglas, 2010; McNamara, 1996; Shohamy, 1994). Indeed, it is necessary for language competence constructs to be firmly embedded in L2 speaking assessment (Chapelle, 1998; Purpura, 2017). As this dissertation is primarily concerned with L2 teachers' oral teacher language competence, there is thus a need for a clear specification on how oral language competence is understood and applied in this particular context. Building on the above considerations of *communicative language ability* (cf. chapter 2.1), Fulcher (2003)'s

definition of oral language competence (i.e. speaking ability) provides the most suitable theoretical basis for this dissertation. In close alignment with Bachman and Palmer (1996), Fulcher (2003) conceptualises speaking ability as being comprised of the following five components (cf. Seong, 2017):

> (1) *language competence* described as phonology, accuracy of syntax, vocabulary and cohesion, and fluency; (2) *textual knowledge* or the understanding of discourse structures such as turn-taking, adjacency pairs, and openings and closings; (3) *pragmatic knowledge* of appropriacy, implicature (doing things with words), and expressing being (defining status and role through speech); (4) *sociolinguistic knowledge* that is situational, topical, and cultural; and finally (5) *strategic capacity* that entails the speakers' use of achievement and avoidance strategies in order to overcome or avoid communication problems. (p. 36, italics in original)

This definition itself highlights the second important factor that needs to be considered when assessing L2 oral language competence and/or performance, namely that it is essential that the test tasks are developed in alignment with the specific features of spoken language. It is only then that reliable test results can be generated that indicate a speaker's ability to speak a language (Arras, 2011; Luoma, 2009). Aside from considering the above five components of speaking ability, there is a range of features that characterise oral language production. A first specific feature constitutes that speech can be *heard*. The audible, vocal sound of speech consciously or subconsciously leads to judgements and interpretations on the speaker *and* the hearer's side, both of whom are naturally prone to a variety of biases that influence their judgements (Luoma, 2009). This is one of the reasons why the sound of speech is a particularly meaningful (and determining) factor in speaking assessment (ibid.). The evaluation of the sound of speech can be approached from a range of perspectives. For example, it can be assessed against criteria of accuracy of pronunciation, with reference to the expressiveness of the speaker's voice, in terms of its comprehensibility or regarding its interactional efficiency (Luoma, 2009). A second distinct feature constitutes that oral language performance follows its own grammar conventions. As opposed to written sentence structures, speaking does not usually include sentences. Instead, speech consists of what Luoma (2009) terms idea units – "short phrases and clauses connected with *and*, *or*, *but* or *that* [that may not be] joined by conjunctions at all but simply spoken next to each other" (ibid. p. 12, emphasis in original). Spoken grammar is much simpler than written grammar because of the real-time nature of speaking that requires speakers to plan, process and produce the (foreign) language. A third

characteristic relates to vocabulary use (Luoma, 2009). Assessment rubrics generally reward rich, complex and precise vocabulary in both speaking and writing assessments. While specific vocabulary can be important in professional contexts or when trying to convey detailed information, generic vocabulary is much more common in spoken interaction. As Sari Luoma (2009) explains,

> [e]ven though [generic words] are not precise, they are fully comprehensible in the speaking situation because they talk about people, things or activities that can be seen or because they are familiar to the speakers. They make spoken communication quick and easy, and few people would find anything strange about this in their mother tongue. (p. 17)

Indeed, generic vocabulary and vague words (e.g., thing, stuff etc.) are important features for the naturalness of speech and for sustaining speech if the speaker cannot think of the word she or he would like to use (ibid.). Other strategies to sustain speech are the use of fillers (ah, kind of, sort of, I don't know), hesitation markers (err, umm), repetition of words or fixed conventional phrases (at the end of the day or all things considered). Just like vague words, frequent repetitions, fillers, hesitation markers or fixed conventional phrases (i.e. lexicalised sentence stems) are often wrongfully punished in speaking assessments. However, using such strategies creates time for a speaker to judge the situation, plan her or his utterance or think of what to say next, and if employed successfully, they should be rewarded (Luoma, 2009). A fourth specific feature of oral language production relates to the integral role of slips and errors. Natural speech frequently contains mispronounced or wrong words or faulty grammatical structures. While such phenomena are normal and less noticed in the speech of a native speaker, they usually receive strong emphasis in L2 speaking assessment to a learner's disadvantage. Finally, further central factors of oral language production constitute its a) ephemerality, meaning that once an utterance has been produced, it vanishes unless recorded, b) intangibility, meaning that it is often vague and inexplicit, or c) (often almost immediate) reciprocity, meaning that speakers take turns to process the demands of speech and produce the text of their speech (Luoma, 2009).

In addition to the need to consider the specific features of oral language production such as those outlined above, understanding of the major role and significant impact of context on language performance is of central importance when it comes to assessing speaking (Chapelle, 1998). Language performance is always context dependent, and a language user needs to use cognitive and metacognitive strategies to activate and employ the relevant knowledge or

competences to successfully complete a given task (Seong, 2017). Chapelle (1998) emphasises that context, strategies, and the interaction of the two, must be considered in the test development process. Purpura (2017) adds that fair, valid and reliable speaking assessments also need to clearly define and address the scope and type of content and *meaning* that is to be measured. These clarifications aid the evaluators to assess the extent to which a test taker's response is *content-responsible*. Aside from the need for language competence constructs to be integrated in L2 speaking assessment, Chapelle (1998) and Purpura (2017) thus highlight the importance of integrating meaningful and relevant content and considering cognitive processes when designing speaking assessments. Finally, spoken language "consists of inherently difficult patterns for humans to attend to" (Port, 2007, p. 362), because of its ephemerality and intangibility, and because synchronous evaluations of spoken language performance are especially prone to biases and subconscious judgements. Consequently, technology may be needed to assist the assessment and rating processes. Indeed, technology has transformed speech into a tangible entity that can now be quantified instrumentally (Isaacs, 2016). For example, video or audio recordings enable to make speech productions (audio-)visually tangible, to generate and digitally store a record of a performance, and to therefore afford speech some permanency (Hewlett & Beck, 2006; Isaacs, 2016). A speaking performance that has been captured by means of technology offers possibilities for scoring or transcription after its live occurrence. For example, a recorded speech production can be rated by any number of human raters who may not have been present at the L2 performance (Isaacs, 2016). Additional affordances of technology-assisted speaking assessments allow, for example, to embed standardised stimuli (e.g., pre-recorded speech elicitation prompts or audio-mediated instructions) to elicit the desired test-taker performance (Isaacs, 2016). Technological assistance opens up new avenues for overcoming some challenges specifically related to speaking assessment. Other challenges to assessing speaking however remain, like for example the need "to reduce a large amount of observational complexity into scores which maintain meaningfulness and interpretability" (Blömeke et al., 2015, p. 9). While technology may moderate this to some extent, it does not provide a silver bullet. The outlined list of characteristics of spoken language are necessary for consideration in speaking assessments. Nevertheless, this list is not conclusive. While all factors are influential in their own right, it is particularly the special nature of spoken grammar and spoken vocabulary that should be of central concern for assessment design (Luoma, 2009). Sari Luoma (2009) consolidates the

elaborations above by suggesting two overall central implications for speaking assessment design:

> Firstly, we must analyse the kind of speaking that we need to assess in a particular assessment context in terms of social and situational needs. Secondly, we must remember that speaking is interactive when we design rating criteria and procedures, and reward examinees when they repeat or mirror other speaker's phrases and structures or develop topics by referring to earlier turns and building on them, because this shows that they know how to work interactively with other speakers. (p. 28)

While these implications and recommendations very much make sense in theory, the practical implementation of oral L2 assessments with the goal to meet as many of the above requirements is a noble, if not almost impossible quest. I will now proceed to discussing ways in which (general and oral) teacher language competence has been assessed including their challenges and limitations. I thereby draw on the elaborations on performance testing, oral performance / competence assessment, communicative competence and profession-related language competence.

2.5.3. Assessing Teachers' Second Language Performance

The increasing (national and international) drive for setting standards for teaching quality causes a surge in the demand for appropriate instruments to assess L2 teacher's language proficiency (Freeman et al., 2009). There are many ways of how this is currently addressed. After a general introduction to LAPPs and performance tests in chapters 2.5.2.1 and 2.5.2.2, I intend to outline the challenges of such approaches based on the specifics of L2 performance assessment in vocational contexts and LAPP settings such as L2 teaching. It builds on the evolving construct of teacher language competence and the mess it causes when attempting to conceptualise it (see also chapter 2.3). Language teaching is considered a highly complex and multifaceted activity which distinguishes itself from other subjects because in the L2 classroom, language is both the medium and the object of instruction. This dual phenomenon implies that L2 teaching relates both the teaching content *and* the teaching process through language (Freeman et al., 2009). Assessing L2 teachers' L2 performance needs to account for this duality. It also needs to ground on a precise definition of the test construct; i.e. it needs to be very clear *what* exactly constitutes the subject of inquiry in order to deduct any possible ways of *how* it can be assessed (cf. chapter 2.2.4; Douglas, 2010; McNamara, 1996; Shohamy, 1994).

However, even recent conceptualisations of teacher language competence are still fuzzy and vague. More precisely narrowing down the construct is a problem that has not yet adequately and satisfactorily been achieved (see chapter 2.3). Since L2 teacher education builds both on academic *and* vocational education, both respective theoretical contexts are applicable. The union of these challenges render capturing and assessing teacher language competence a highly rich, complex and volatile enterprise (Freeman et al., 2009). The common and misleading assumption that "tests of general language proficiency […] classify the workplace 'readiness' of people's language abilities" (Knoch & Macqueen, 2020, see chapter 2.5.2.1), for instance, is made explicit as a prime example in a survey conducted with Swiss universities of teacher education (Hunkeler et al., 2009). The results show that, at the time of the survey, most universities of teacher education in Switzerland accepted common international language diplomas (ILD) to certify the language competences of graduating L2 teachers (Hunkeler et al., 2009). Some of the reasons mentioned that might explain the popularity of ILD are that they are considered objective, comparable, professionally administered and widely recognised. The perceived benefits thus largely outweigh the costs and efforts connected to developing and administering internally organised LAPP tests to certify profession-related language competences (ibid. p. 15). A prominent yet conflicting finding of the survey is that, despite the common acceptance of ILD, the required proficiency standards for graduating L2 teachers vary across institutions. There are also considerable differences with reference to the implementation of additional assessments of profession-related, as opposed to generic, language competences (Hunkeler et al., 2009). While some Swiss institutions "only" require an ILD to certify the required L2 competences, others conduct their own internal assessments on profession-related aspects of the L2. Just like with the overall handling of L2 certification, there is stark variation regarding the additional internal assessments on profession-related language competences. Most of these tests are developed within the respective institutions; some replicate the structure and contents of ILD while others represent their own individual assessment types, which again vary in terms of their orientation on frameworks of reference. While some rely on the CEFR, others align with the institution's curriculum (Hunkeler et al., 2009). Practical exams conducted in internships constitute yet another form (Hunkeler et al., 2009). The results thus present a highly heterogeneous picture regarding efforts and approaches to attest pre-service teachers' attainment of L2 competences in Switzerland. In addition, ILD focus on testing generic language competences required mostly for everyday use (ibid.). They hence fail to specifically address and evaluate profession-related language competences or establish a connection to the

relevant professional domain, especially for the purpose of a certification of such skills within L2 teacher education programmes (Bleichenbacher et al., 2014). This became apparent during a systematic comparison of selected ILD with the PRLCP (Bleichenbacher et al., 2014). Even though this comparison was conducted as a pilot study and findings cannot be seen as conclusive, the insights gained are nevertheless intriguing. The results indicate that ILD cover communicative competences related to the PRLCP – and thus the teaching profession – more comprehensively the higher the CEFR level they certify (see Table 2). There are also stark differences between ILD of different languages. For example, ILD that certify German L2 competences tend to cover certain profession-related language aspects more comprehensively than ILD that certify French or English. The most striking finding however is that there is great variation related to how comprehensively ILD cover profession-related language competences described across the different AoAs of the PRLCP. For example, a specific set of competences that are particularly underrepresented in, and at numerous instances even absent from common ILD descriptors, are skills related to *evaluating, giving feedback and advising* (AoA 3). Row *HF 3 Total* of the below Table 2 highlights this. The first three columns on the left (*HF3*, i.e. AoA 3, *Sprache*, i.e. language, and *Kompetenzprofile*, i.e. PRLCP) indicate the amount of communicative tasks and descriptors that are represented in AoA 3 per mode (reading, speaking or writing), language, and school level (*Primar*, i.e. primary school, and *Sek. I*, i.e. lower secondary school). For example, the PRLCP contain 7 speaking descriptors in AoA 3 that are relevant for lower secondary school teachers. The last three columns indicate how many of the available AoA 3 descriptors are represented in common ILD, separated according to the CEFR levels the ILD certify. For instance, ILD that certify English oral competences at CEFR level C2 only cover 1/7 speaking descriptors of AoA 3 at *Sek I* level:

	Sprache	Kompetenzprofile		Alle ISD für eine Fremdsprache					
				B2		C1		C2	
		Primar	Sek. I	Primar	Sek. I	Primar	Sek. I	Primar	Sek. I
HF3: Beurteilen, Rückmeldungen geben und beraten									
Lesen	Franz.	1	1	0/ 1	0/ 1	0/ 1	0/ 1	0/ 1	0/ 1
Lesen	Engl.	1	1	0/ 1	0/ 1	0/ 1	0/ 1	0/ 1	0/ 1
Lesen	Deut.	1	1	1/ 1	1/ 1	1/ 1	1/ 1	1/ 1	1/ 1
Sprechen	Franz.	5	7	0/ 5	0/ 7	0/ 5	0/ 7	0/ 5	0/ 7
Sprechen	Engl.	5	7	0/ 5	0/ 7	1/ 5	1/ 7	1/ 5	1/ 7
Sprechen	Deut.	5	7	0/ 5	0/ 7	0/ 5	0/ 7	0/ 5	0/ 7
Schreiben	Franz.	3	3	0/ 3	0/ 3	0/ 3	0/ 3	0/ 3	0/ 3
Schreiben	Engl.	3	3	1/ 3	1/ 3	1/ 3	1/ 3	1/ 3	1/ 3
Schreiben	Deut.	3	3	1/ 3	1/ 3	1/ 3	1/ 3	1/ 3	1/ 3
HF3 Total	Franz.	10	12	0%	0%	0%	0%	0%	0%
HF3 Total	Engl.	10	12	10%	8%	20%	17%	20%	17%
HF3 Total	Deut.	10	12	20%	17%	20%	17%	20%	17%

Other criticism related to employing ILD to certify teachers' L2 competences are connected to findings suggesting that ILD overemphasise grammatical knowledge over oral proficiency. Indeed, speaking competences are not considered satisfactorily relative to the requirements of the teaching profession. In an analysis of the *Cambridge Advanced Examination (CAE)*, for example, Bader-Lehmann (2007) found that a Cambridge exam can be passed despite unsatisfactory oral language competences (p. 244-245). This leads to the concern that even if pre-

Table 2 : ILD coverage of PRLCP descriptors (excerpt) (cf. Bleichenbacher et al., 2014)

service teachers achieve a C1 level in a CAE, their speaking competences might not meet the required standards of L2 teaching (Bader-Lehmann, 2007). Finally, teacher qualification often serves to measure teacher content knowledge – a construct that is highly challenging to measure (Loder-Büchel, 2014). However, such tests are not differentiated enough to provide in-depth information about a teacher's profession-related language competences. It is important to note that teacher certification or ILD neither provide evidence for a teacher's expertise or teaching effectiveness nor establish connections to their learner's performance. In sum, there is evidence that ILD tend to assess mostly generic language competences and are thus not a valid instrument to use when it comes to assessing teacher language competence. This deficit consequently implies that pre-service teachers lack the opportunity to provide evidence of their proficiency or *knowledge-in-action* in several central aspects of their profession-related language skills. It also means that L2 teacher education programmes lack a means to fully and satisfactorily assess and certify teacher language competence, and that the efforts made to find ways of doing so are highly heterogeneous across Switzerland (Hunkeler et al., 2009). Hunkeler et al. (2009) underline the increasing need for assessment systems to evaluate and test profession-related language competences, especially when considering that the acquisition of these competences is a requirement for the completion of the teacher education programmes. Much in line with the purpose of the PRLCP and the recommendations of the Swiss Conference of Cantonal Ministers of Education (EDK)[14] and *swissuniversities*, discussions around the development of suitable such assessment systems are becoming increasingly more prominent. This desideratum is not restricted to the Swiss context. Primarily in the context of *English-for-teaching* but also beyond,

[14] The EDK is an inter-cantonal political body responsible for coordinating Swiss national education and culture policy. It is comprised of a commission of educators and cantonal board of education directors. By setting guidelines and issuing relevant recommendations for standards in education on the primary, secondary and tertiary levels, the EDK aims to ensure high quality, equity, permeability and mobility within the Swiss education system. Consult http://www.edk.ch for more information.

Elder and Kim (2014) provide an (illustrative) overview of the different types of tests that are commonly used to assess teacher language competence:

- General proficiency tests used for teacher certification
    - Assessments for L2 teachers (e.g., the American Council on the Teaching of Foreign Languages Oral Proficiency Interview: ACTFL OPI),
    - Standardised tests of academic English proficiency for non-native teachers in English-medium education contexts (e.g., IELTS; TOEFL)
- Tests specifically targeting teacher language proficiency
    - English for specific purposes (ESP) tests for non-native teachers in English-medium teaching contexts
        - Professional English Assessment for Teachers (PEAT)
        - Taped Evaluation of Assistants' Classroom Handling (TEACH)
        - Classroom Language Assessment Schedule (CLAsS)
- LSP tests of language teacher proficiency
    - Language Proficiency Test for Teachers (LPTT)
    - Teste de Proficiência Oralem Língua Inglesa (TEPOLI)
    - The Language Proficiency Assessment for Teachers of English (LPATE)
- LSP tests for teachers of bilingual education
    - Arizona Classroom Teacher Spanish Proficiency Exam (ACTSPE)
    - National Māori Language Proficiency Examinations (NMLPE) (ibid.)

In their review, Elder and Kim (2014) critically discuss these tests and question whether they can guarantee reliable results on teachers' profession-related L2 competences that certify a person's adequacy for entering the profession. According to Elder and Kim,

> while some general proficiency measures containing a performance-based speaking component may suffice for screening purposes, they do not guarantee classroom readiness and may have limited utility as diagnostic tools to support the teaching and learning of communication skills relevant to the teaching domain. (p. 5)

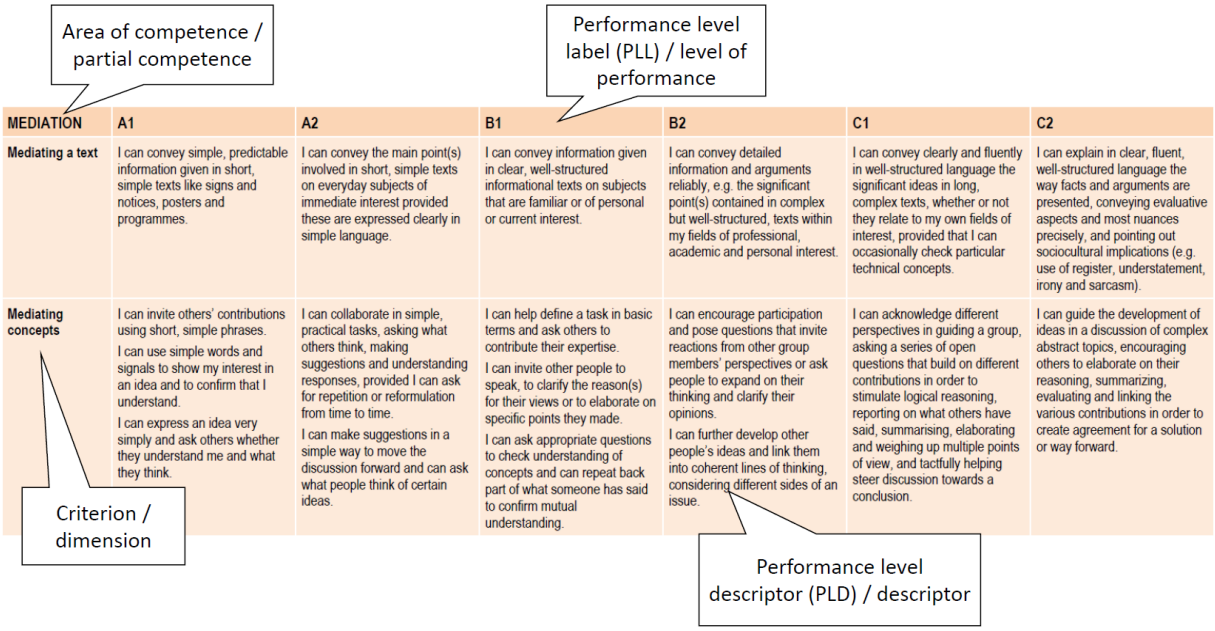
A further problem related to such tests is that they often invoke an idealised notion of native-speaker competence as a benchmark (see chapter 2.3). Such an approach is unhelpful, it overemphasises and overvalues a teacher's subject knowledge and likely under-represents the teacher language competence construct (see chapter 2.3). The multifaceted nature of the construct has of yet prevented to reach clarity on "how much proficiency is enough for effective teaching performance", resulting in a lack of "explicit justification or empirical evidence for existing minimum thresholds or relative weightings given to general versus profession-specific abilities" (ibid. p. 14). In addition, adequately eliciting and capturing the relevant language performances in a language test is challenging (Elder & Kim, 2014). Apart from the "common" challenges of performance tests (see chapter 2.5.2.2), assessing teacher language competence needs to confront the inherent difficulty to capture "[t]he complex array of language functions and discourse strategies involved in interacting appropriately with learners who may have limited command of the TL" (ibid. p. 14). It is not only the validity of the existing tests that has been challenged, but also the reliability (Chalhoub-Deville & Fulcher, 2003; Elder, 2001; Salaberry, 2000). In addition, ethical considerations and arguments with reference to their fairness and equity based on the high-stakes nature of certain tests have been voiced (Burke, 2013; Norris, 2013). Elder and Kim (2014) and Burke (2015) address such issues by explicitly calling for more evidence-based, reliable, fair, independent and standardised language tests:

> It is imperative that third-party expert reviewers, not funded or chosen by accrediting bodies or affiliates, be allowed and encouraged to engage in empirical research to examine the reliability, validity, and ethicality of general language proficiency and academic proficiency tests for teachers being used to determine the future of individuals in terms of their licensure, certification, employment, promotion, admission, and graduation. (Burke, 2015, p. 5)

Finally, the above considerations need to reflect the current pedagogical, administrative and technological practices in which the assessments are situated (Bearman et al., 2020). In the early 2020s, this means that they must "align with a digitally enabled world with rapidly expanding information and an increasingly dynamic view of knowledge" (ibid. p. 7). After this LSP lens on language testing, the subsequent chapter zooms back out to general aspects of assessing L2 oral language competence and outlines different ways of how it can be scored.

2.5.4. Scoring (Teachers') Second Language Performance

Evaluation and scoring practices constitute a central and discriminating aspect of language tests. Scoring practices do not only establish the link between a performance and a proficiency claim, but also play a pivotal role in conceptualisations of validity (Kane, 2013; Knoch et al., 2021). There are a variety of scoring methods, which, again, vary depending on the test construct and test purpose. While in traditional fixed-response assessment scores are "derived directly from the instrument, which (for each item) offers the candidate a number of choices, only one of which is correct" (McNamara, 1996, p. 120), in communicative language tests or performance tests the elicited language performances are generally judged against relevant assessment criteria (McNamara, 1996). Language performance can generally be scored by means of automated, computerised scoring, or by means of human raters using rating scales or checklists (Council of Europe, 2001). Checklist-mediated rating involves, as implied by its name, a list of aspects that are considered relevant for a particular language course or module. When employing checklists to score a language performance, the emphasis lies on indicating how much of the course content or how many of the learning goals a test taker has successfully achieved (Council of Europe, 2001). Rating-scale mediated scoring, in contrast, involves "judging that a person is at a particular level or band on a scale made up of a number of such levels or bands" (Council of Europe, 2001, p. 189). Both approaches include human raters. Thus, this type of assessment is categorised as *rater-mediated assessment*. Rating scales are a collection of relevant criteria or dimensions (henceforth: criteria) that contain Performance Level Labels (PLLs, terms used to label performance categories) and Performance Level Descriptions (PLDs, verbal elaborations of the knowledge, skills, or attributes of test takers within a performance level, cf. Cizek & Bunch, 2007). The following example from a self-assessment scoring rubric extracted from the CEFR-CV (Council of Europe, 2018, p. 170) illustrates common components of a rating scale:



MEDIATION	A1	A2	B1	B2	C1	C2
Mediating a text	I can convey simple, predictable information given in short, simple texts like signs and notices, posters and programmes.	I can convey the main point(s) involved in short, simple texts on everyday subjects of immediate interest provided these are expressed clearly in simple language.	I can convey information given in clear, well-structured informational texts on subjects that are familiar or of personal or current interest.	I can convey detailed information and arguments reliably, e.g. the significant point(s) contained in complex but well-structured, texts within my fields of professional, academic and personal interest.	I can convey clearly and fluently in well-structured language the significant ideas in long, complex texts, whether or not they relate to my own fields of interest, provided that I can occasionally check particular technical concepts.	I can explain in clear, fluent, well-structured language the way facts and arguments are presented, conveying evaluative aspects and most nuances precisely, and pointing out sociocultural implications (e.g. use of register, understatement, irony and sarcasm).
Mediating concepts	I can invite others' contributions using short, simple phrases. I can use simple words and signals to show my interest in an idea and to confirm that I understand. I can express an idea very simply and ask others whether they understand me and what they think.	I can collaborate in simple, practical tasks, asking what others think, making suggestions and understanding responses, provided I can ask for repetition or reformulation from time to time. I can make suggestions in a simple way to move the discussion forward and can ask what people think of certain ideas.	I can help define a task in basic terms and ask others to contribute their expertise. I can invite other people to speak, to clarify the reason(s) for their views or to elaborate on specific points they made. I can ask appropriate questions to check understanding of concepts and can repeat back part of what someone has said to confirm mutual understanding.	I can encourage participation and pose questions that invite reactions from other group members' perspectives or ask people to expand on their thinking and clarify their opinions. I can further develop other people's ideas and link them into coherent lines of thinking, considering different sides of an issue.	I can acknowledge different perspectives in guiding a group, asking a series of open questions that build on different contributions in order to stimulate logical reasoning, reporting on what others have said, summarising, elaborating and weighing up multiple points of view, and tactfully helping steer discussion towards a conclusion.	I can guide the development of ideas in a discussion of complex abstract topics, encouraging others to elaborate on their reasoning, summarizing, evaluating and linking the various contributions in order to create agreement for a solution or way forward.

Figure 11 : Illustrative example of a (self-assessment) rating scale

Raters use rating scales and criteria to make judgements about test takers' performances by assigning them to a PLD (henceforth: descriptor or PLD). This combination of "agents" (raters, rating scales, rating criteria, PLD) introduces a new type of interaction between a number of factors. As McNamara (1996) points out, this interaction primarily concerns the relationship

> between the rater and the scale; this interaction mediates the scoring of the performance. The rater-scale interaction resembles the subject-instrument interaction in that the rater-scale interaction is like a 'test' of the raters (and the scale) in the way that the subject-instrument interaction is a test of the subjects (and of the instrument). Just as we have always sought information on the instrument and the subject, so we should seek information on the scale and the raters […]. (p. 121)

Rating criteria guide the rating process by making "implicit reference to a psychological construct or constructs, which then emerge as the object of measurement" (McNamara, 1996, p. 19). Rating scales are not only central for the rating process per se, but also for the communication of test data (McNamara, 1996). Because they form part of the test construct definition, they are central to any communicative performance test (Luoma, 2009). With their significant role in any assessment procedure, the design of rating scales can have a

> direct effect on score generalizability (how broadly the score can be generalized beyond the test-taking situation), the precision of the predictions that can be made about test takers, as well as on the reliability of the scores. (Knoch et al., 2021, p. 3)

Thus, their determination is central to the test development process and the test's overall validity, reliability and objectivity. Rating scales can be distinguished according to their orientation (user-, assessor- or constructor-orientation, cf. Alderson, 1991) or according to the type of descriptor (i.e., descriptors of *communicative abilities* or descriptors of *aspects of proficiency* related to particular competences, cf. Council of Europe, 2001). In performance assessment in particular, statements that define what test takers can do (*can-do descriptors*) rather than what test takers *know* are more conducive to achieving a reliable judgement (McNamara, 1996). Because of the crucial role of rating scales and PLDs in communicative and performance assessment, developing rating scales requires particular rigour. General best practice principles include that descriptors be formulated *positively*, have a high degree of concreteness (in the sense of *precision* or *definiteness*) and *clarity* (in the sense of transparency), and are formulated precisely and concisely (*brevity*). Further, they need to be *independent* from one another as independent and integral criteria statements (Council of Europe, 2001), and they must be *meaningful to raters* and *relate meaningfully to real-world language use* (Knoch et al., 2021). Moreover, the consistency with which teachers and learners can interpret descriptors of communicative abilities is enhanced if they describe both *what* a learner can do and *how well* they do what they can do (Council of Europe, 2001). It is evident that the development of valid and reliable rating scales is a complex endeavor, especially when it comes to rubrics for performance tests. In his PhD thesis, Brian North (2000) describes this challenge as "trying to describe complex phenomena in a small number of words based on an incomplete theory" with no empirical foundation based on the analysis of actual performance data. In this context, Fulcher et al. (2011) partly disagree when differentiating between a measurement-driven and performance-driven approach to developing rating scales. The former usually involves intuition-driven expert opinions (e.g., any rating scales derived from the CEFR descriptors). Hence, the measurement-driven approach results in scales that "express the developer's understanding of how good performances differ from weak ones" (North, 2000, p. 59). In contrast, the performance-driven approach bases the scale development on actual performance data or corpora. Because of their evidence-based constructions, Fulcher et al. (2011) argue that well-designed performance-driven scales reflect real-world language use better because they allow for the imperfections, hesitations and false starts that are characteristic of (spoken) language use (see chapter 2.5.2.3). This dichotomous measurement- versus performance-driven logic has been criticised for not corresponding to current and actual practice. Indeed, rating scale development may combine a range of methods including the incorporation of expert

opinions, insights from performance data analyses or other *a priori* or *a posteriori* qualitative and quantitative methods (cf. Knoch et al., 2021 for an overview of sources that have influenced rating scale development processes and the rating scale construct in published research studies on rating scale construction). Knoch et al. (2021) emphasise that any scale development procedure includes a broad range of complex decisions and maintains that the development process needs to start with carefully considering the test purpose and the score use, and ensuring that the test construct aligns with the intended score use. It is at this stage in particular where the following two questions are often conflated with performance assessments' test purposes: Are the criteria chosen to assess language as the *medium* of a performance, i.e. language as part of a communicative act to assess a test taker's capacity to perform a task? Or are they chosen to assess a second language performance elicited through a task that simulates a real-world situation (McNamara, 1996)? This conflation becomes explicit in what several authors consider a major shortcoming of LAPP performance tests (Douglas, 2000; Knoch & Macqueen, 2020; McNamara, 1996). Even though they are most often developed based on extensive needs analyses and should therefore provide a strong link to the TLU (much like the PRLCP), their assessment criteria are mostly linguistic in nature and fail to reflect the views of domain experts on what constitutes a strong performance (Knoch & Macqueen, 2020). Thus, such assessment criteria lack "indigenous criteria", namely the values held by domain insiders about what constitutes effective spoken or written communication in that particular TLU domain (Jacoby, 1998). As a result, such tests can at best be *weak* performance tests (see chapter 2.5.2.2). In any case, rating scales are a necessary and unavoidable evil of test development, which, just like the overall language test, need to meet quality criteria that ensure reliability, validity and objectivity of the evaluation method, the evaluation criteria and the evaluators. In the following subsections, I will zoom in further and explain two common types of scoring. Other forms of scoring are not discussed; however, the CEFR (Council of Europe, 2001) provides a succinct yet comprehensive overview for further reference.

2.5.4.1. Holistic Rating

Global assessment (i.e. *holistic assessment*) involves making a global, synthetic judgement against established rating criteria where different aspects are weighted intuitively by the rater (Council of Europe, 2001). Holistic scales allow raters to award a general, single and overall score to a test taker's performance on a whole test. They are thus flexible and insofar global as they allow for the evaluation of the entire performance based on increasing levels of overall achievement. Holistic rating scales are developed based on a more traditional approach where

language ability is understood as a "single unitary ability" (Bachman & Palmer, 1996). Because global scales are so broad, they naturally contain multiple "hidden" components of language ability (ibid. p. 339). Diverging and individual manifestations of different dimensions as well as different combinations of those modes on different levels are thus not uncommon. The practicality and efficiency of holistic scoring, the ease of reporting (Davies et al., 1999) and the fact that holistic rating processes place lower cognitive demand on the raters (Xi, 2007) are some advantages of this scoring procedure. While this approach may ease the process of reaching consensus between evaluators, there are also a number of problems associated with it. Bachman and Palmer (1996) point out, for instance, that global scales do not display what exactly it is that a score discloses. Indeed, it remains unclear whether a score reflects for instance multiple areas of knowledge (e.g., vocabulary, pronunciation, accuracy, cohesion, etc.) or multiple modalities (e.g., production, reception, interaction or mediation). These confounding aspects of language use thus obscure the true meaning of raters' scale interpretations and awarded scores. If a test seeks to reveal more fine-grained information on (speaking) performances, holistic rating procedures are not appropriate (ibid.). In addition, the globalism of the scales often make it difficult for raters to assign a production to a specific level. For example, if a language production contains highly complex grammatical structures and frequent errors, a rater has to decide which of these aspects takes precedence when assigning it to a level of the criterion "language use" (Bachman & Palmer, 1996). Moreover, individuals may consciously or subconsciously weigh the hidden components differently (Bachman & Palmer, 1996). In this case, a given rater A may prioritise complexity of structures over accuracy and a given rater B vice versa, but both raters may assign the same level (ibid.). These cognitive processes remain concealed and thus no precise judgement about a test taker's language ability is possible. However, global scales carry the potential for raters to largely focus on what candidates can do rather than on where they display difficulties (Bacha, 2001). This is in line with the goal of competence-oriented language testing to reveal what a test taker can accomplish with her or his current language skills as opposed to what she or he cannot (Council of Europe, 2001). Regardless of whether raters focus on candidates' strengths or weaknesses (or both), holistic scoring does not allow for the manifestation of ability in specific areas to be made explicit. Finally, global scales are constructed based on the underlying assumption that all areas of language competence develop uniformly, linearly and homogenously (Kroll, 1990). From a second language acquisition perspective, however, this approach is questionable (L. Bachman & A. Palmer, 1996; Europe, 2001; Khabbazbashi & Galaczi, 2020).

2.5.4.2. Analytic Rating

In contrast to global rating, analytic assessment involves evaluating distinct aspects of L2 performance separately (Council of Europe, 2001). Analytic scoring entails raters assigning a separate score to a variety of explicitly defined evaluation criteria which are related to different aspects of performance (Khabbazbashi & Galaczi, 2020; Xi, 2007). The basic assumptions underlying analytic scoring ground on multidimensional language competence models and allow for the discrete evaluation of individual language-specific, content-related and communicative performances (Grum, 2012). Thus, a multi-dimensional analytic rating scale reflects the complexity of language use more accurately (Khabbazbashi & Galaczi, 2020). Indeed, as there is broad consensus that language acquisition is not a linear process, the language ability dimensions need to be evaluated separately (Bachman & Palmer, 1996; Council of Europe, 2001; Grum, 2012). In contrast to holistic scales, analytic scales ensure a more systematic evaluation process (Khabbazbashi & Galaczi, 2020) that makes explicit the components included in the construct at hand (Bachman & Palmer, 1996). This explicitness enables more transparency of how raters weigh the different components (Khabbazbashi & Galaczi, 2020) and more control of rating behaviour (e.g., through training raters in the application of the rating scales) (Bachman & Palmer, 1996). As analytic rating scales reflect the different components of language ability and the criteria and dimensions more transparently, the raters are also more likely to have clarity about what to focus on when they score performances (Khabbazbashi & Galaczi, 2020). Because of the "profile" of the language ability components that are provided in an analytic rating scale, the relative strengths and weaknesses of test takers can be identified and reported more easily (Bachman & Palmer, 1996). Such profiles allow for a more accurate picture of the different developmental stages a learner goes through when learning an L2. Such information can be used for diagnostic purposes and for helping learners with uneven profiles to progress through offering more targeted support (Khabbazbashi & Galaczi, 2020). Another advantage of analytic scales is that they provide a more accurate (albeit still not fully transparent) reflection of raters' behaviour (Bachman & Palmer, 1996). Despite the benefits of analytic scoring, the complexity of the process places higher cognitive demands on raters (Khabbazbashi & Galaczi, 2020; Xi, 2007). This aspect alone makes the scoring process highly vulnerable to rater effects. To mitigate the cognitive load, strategies need to be explicitly taught and implemented. Such strategies may involve, for example, limiting the amount of performances to be rated per "rating episode" (current best practice principles suggest a maximum of five performances at a time; cf. Council of Europe,

2001). Also, limiting the amount of criteria to be assessed and prescribing how often and in which intervals raters need to take breaks are common strategies to reduce cognitive load (Grum, 2012). The CEFR specifies "more than 4 or 5 categories starts to cause cognitive overload and that 7 categories is psychologically an upper limit" (Council of Europe, 2001). In addition, rating criteria need to correspond to the best practice principles outlined above (see chapter 2.5.4) and be explicitly defined, clear-cut and comprehensive, otherwise raters may experience difficulties distinguishing between the criteria as well as the dimensions (Xi, 2007). Based on these considerations, analytic scoring rubrics seem more suitable for assessing (oral) L2 teacher language competence than holistic rubrics. The following subchapter will introduce an analytic scoring rubric that has been developed with the particular purpose to do just that: assess profession-related language competences.

2.5.4.3. PRLC-R

As outlined in chapter 2.5.4, developing valid and reliable assessment rubrics for performance tests is a particular challenge. One of the most common and major shortcomings of LAPP performance tests is that the assessment criteria are often linguistic in nature and fail to reflect both the TLU as well as the views of domain experts on what constitutes a strong performance (Douglas, 2000; Knoch & Macqueen, 2020; McNamara, 1996). The lack of indigenous criteria also concerns the PRLC-R, which, as a central tool for describing and assessing profession-related language competences, lie at the heart of the present dissertation. The PRLC-R is an analytic assessment rubric that roots in the PRLCP and contains scales with PLDs illustrating relevant language-teacher-related competences. The PRLC-R were developed with the aim to create an assessment rubric closely aligned with the PRLCP that allows raters to make valid judgements about an L2 teacher's profession-related language competences. The rubric contains scales for the following skills: 1) *general: task completion,* 2) *listening,* 3) *reading,* 4) *qualitative characteristics of speaking* including (4a) *spoken language production* and (4b) *spoken language interaction,* and 5) *qualitative characteristics of writing.* The scale *qualitative characteristics of speaking* contains the following criteria: *task completion, vocabulary, accuracy, pronunciation, fluency, cohesion and coherence*, and *addressee-specificity* (see also chapter 4.2). The assessment criteria (with exception of the dimensions *task completion* and, to some degree, *addressee-specificity*) solely encompass linguistic factors. *Addressee-specificity* refers to a language user's ability to adapt their expression to the proficiency level and needs of the target group in order to ensure understanding (and, ideally, induce learning). *Addressee-specificity* can be interpreted as indigenous and, despite being included as an equally weighted

criterion in the rubric, it seems to address a markedly different dimension on a different level than the other criteria. Although the developers of the PRLC-R argue that the outlined linguistic criteria are *profession-related* and thus represent *indigenous* criteria to provide a profession-related lens on performance, they in actuality only construe a "face resemblance", a superficial impression of profession-relatedness. The PRLC-R thus allow for an interpretation of task-elicited performances in linguistic terms only. With performance tests that build on assessment criteria like these, their focus can only lie on the *language performance* itself rather than the performance of the *task* (McNamara, 1996; Messick, 1994). Thus, any (profession-related performance) test that employs the PRLC-R would be considered a *weak* performance test in the sense of McNamara's (1996) classification (see chapter 2.5.2.2). This means that a respective test purpose would be "to elicit a language sample so that second language proficiency, and perhaps additionally qualities of the execution of the performance, may be assessed" (McNamara, 1996, p. 44). In addition, the PRLC-R reflects some central issues of the construct of teacher language competence that may be difficult, if not impossible, to overcome. For example, a teacher's most elaborate vocabulary cannot be scored highly if it does not help overcome the communicative challenge in context because it will likely not be understood by the target group (i.e. lower-secondary school students). It is thus fair to assume that the common aspects of language quality such as (high) complexity (of vocabulary or structures) and (high) fluency in oral language productions may conflict with the indigenous criterion *addressee-specificity*. This may be particularly problematic in tasks in which the language ability of the addressee is to a large extent inferior to that of the teacher. Indeed, very complex and highly fluent language of high articulation rate are likely to be overwhelming instead of "excellent" for many lower-secondary school students striving to learn an L2 (Bleichenbacher et al., 2019). In sum, the common "faster, higher, stronger" – in other words "complex, accurate, fluent" – paradigm may not always be applicable (ibid.). The fact that this notion may however be relevant – for example if the communication content is of low complexity and the addressee's language skills are highly advanced, e.g., in the case of a bilingual learner – shows that the appropriate application of the PRLC-R is likely to be highly complex. The principles are conflicting in many ways and to be able to use the assessment rubric to release its full potential requires high sophistication on the raters' part. To be able to better understand the potential challenges connected to the application of the PRLC-R, the next section outlines how and why the reliability and validity of judgements by human raters may be threatened.

2.5.4.4. Reliability of Human Raters

Rater-mediated performance assessments involve human raters who judge elicited language performances and make inferences about the test takers' underlying language ability. When raters make such judgements, it is usually a complex undertaking that involves subjective interpretation. Such judgement processes are always subject to disagreement and prone to a variety of errors or influences of error (Eckes, 2011; McNamara, 1996). However, ensuring reliable and objective expert ratings and guaranteeing sufficiently high psychometric quality is an essential component of the quality assurance of any rater-mediated performance test, especially in high-stakes, high-impact tests that carry gate-keeping functions (Eckes, 2005). There are a variety of methods to identify and quantify the extent of agreement between human raters in order to reduce the disagreement to acceptable levels (McNamara, 1996), and to improve the production of "fair measures of the ability of test candidates in performance assessment settings" (McNamara, 1996, p. 117). One of the most important measures is to ensure that the scoring criteria and PLDs are carefully, clearly and precisely formulated (see chapter 2.5.4). Another common measure constitutes the provision of a precise rating manual with benchmark exemplars of performance and examples of the application of the assessment rubric. Aside from providing raters with guiding materials, three methodical steps are traditionally involved to determine and improve the reliability of rating procedures: repeated double ratings of the same tasks by independent raters, the evidence of satisfactory reliability of the ratings, and periodical rater trainings (Eckes, 2011). Double ratings allow for the two main reliability measures in human-rater scored test tasks to be calculated (Douglas, 2010): *interrater reliability* (often also referred to as *intercoder reliability* or *interrater agreement*) and *intrarater reliability* (Eckes, 2011). *Interrater reliability* reports on the degree to which two independent raters score the same performance identically. In contrast, *intrarater reliability* provides information related to the degree (internal consistency) to which a rater scores the same performance identically when scoring it a second time (Douglas, 2010). The extent to which these reliability measures are satisfactory can be determined by calculating a variety of different reliability coefficients. Tinsley and Weiss (1975, 2000) differentiate between consensus and consistency measures (Stemler & Tsai, 2008). While consensus measures represent the degree of absolute agreement between raters, consistency measures assess the degree in which the rated tasks are in agreement with one another (Eckes, 2011). To measure both types of reliability there are several methods that need to be selected based on the particular type of information that needs to be obtained. As Wirtz and Caspar (2002) postulate,

"Koeffizienten bilden jeweils eine bestimmte Eigenschaft der erhobenen Datenstrukturen ab und spiegeln somit jeweils unterschiedliche Informationen wider" (p. 23). Hence it is common to simultaneously report on a variety of different coefficients to gain a more comprehensive understanding of the data structure (Wirtz & Caspar, 2002). The interpretation of computed reliability coefficients needs to be treated with caution because a high agreement between raters does not exclude the possibility that all raters subconsciously rated with the same error (Eckes, 2011). Indeed, original ratings given by raters (i.e. raw scores) have a high potential to be misleading and can never be taken as a completely reliable source to determine a test taker's ability (McNamara, 1996). One way of mitigating this is by repeatedly conducting extensive and rigorous rater trainings – the final commonly applied method to improve the reliability of human ratings. Rater trainings aid expert raters to gain a mutual understanding of the rating scales and reach consensus when it comes to scoring tasks at different competence levels or PLDs. Such training procedures (including the provision of precise and distinct assessment criteria) serve to counteract human raters' natural tendency to freely associate deficiently articulated assessment scales with their own independent ideas, to weigh performances according to their individual measures, or to subconsciously apply non-existent assessment criteria (Eckes, 2011; Grum, 2012). It is only by rigorous rater trainings that reliable and consistent ratings can be promoted, which can then serve as a basis for making reliable inferences about a test taker's underlying language ability. However, while there is evidence that rigorous rater trainings are effective in terms of training raters to be more self-consistent and in terms of reducing extreme differences (e.g., identifying and excluding outliers), perfect agreement is virtually impossible to attain (Douglas, 2010; McNamara, 1996). Stable findings indicate that raters employ their own unique perceptions that are almost immune to alteration through training (Lunz & Stahl, 1990). Individual rater characteristics thus naturally and very typically lead to rater variability. Rater variability is so extensive that it is crucial for this aspect to be considered and compensated for. Aside from rater characteristics, scale characteristics and test task variation may also have large effects on ratings. While the first are generally referred to as *rater effects,* the latter two are known as *interaction effects* (McNamara, 1996). *Interaction effects* include the severity or leniency with which a rater judges a performance, which may also differ in relation to candidates (e.g., rater biases related to gender, accent, age, etc.), or in relation to particular test items. While the former is an instance of a *rater-candidate* interaction, the latter is an instance of a *rater-item* interaction (McNamara, 1996). *Rater effects* are systematic, ubiquitous, perennial and manifold, and thus difficult to control let alone eliminate

(Eckes, 2005). Such effects are also irrelevant of the test constructs and hence threaten the validity of the assessment procedure (Bachman, 2004; Eckes, 2005; Messick, 1995; Weir, 2005). *Rater effects* commonly include the trait that raters differ from one another in terms of how they interpret, interact with and apply the rating scale they are using, like for instance by assigning differently sized intervals between the performance levels of a rating scale (Eckes, 2005). For instance, a given rater A may interpret the distance between performance level 1 and performance level 2 as much larger than between performance level 2 and 3, whereas a given rater B may interpret the intervals between 1 and 2 and 3 more narrowly. Furthermore, raters with a so-called *central tendency* avoid using the ends of the rating scale. They will therefore differ from raters who avoid the middle of the scale and instead have a tendency to see more stark differences between test takers (McNamara, 1996). Raters may also differ in terms of how consistently they rate, as in how uniform their scoring pattern is. This source of variation is also referred to as the extent of the *random error* in a given rater's judgements (McNamara, 1996). As it is very difficult to predict and compensate for harsh inconsistency, severely inconsistent raters need to either be retrained or excluded from the rating process and data analyses (McNamara, 1996). Finally, aspects such as assessment conditions (e.g., intelligibility of audio-recorded responses), halo effects, raters' educational, professional or linguistic (L2) background, the time of the day the ratings are conducted, the amount of ratings completed per session, the cognitive load experienced, the order in which performances are rated etc. additionally influence individual ratings. For example, there is evidence that raters may favour certain language variations (i.e. "accents") over others, indicating raters' L1 background as a potential source of rater bias (Winke et al., 2013). With reference to the order of test responses according to which a rater conducts her or his evaluations, research findings suggest that a low-proficiency test response may influence the rating of a subsequent response of higher proficiency. This is partly attributed to a raters' tendency to subconsciously judge the latter sample more highly than it actually is because of the stark contrast to the previous sample. Such phenomena need to be interpreted based on the notion that raters' judgments are fundamentally relative in nature rather than a pure, error-free application of an absolute rule (Davis, 2012). Despite the difficulty to control rater variability and the problems related to the traditional approaches of compensating for judgemental errors, it is important that the characteristics of raters and rater effects, and the characteristics of tasks and assessment settings are determined and dealt with accordingly to ensure that fair estimates about test takers' abilities can be derived (McNamara, 1996). While the traditional approaches to minimising rater variability are

important and valuable in terms of increasing the validity, objectivity and reliability of a test and ratings as well as in terms of "creating the conditions for an orderly measurement process" (McNamara, 1996), they can only reduce but not fully eliminate rater variation. Indeed, Eckes argues that rater trainings, repeated ratings, and interrater reliability coefficients as traditional approaches to control judgemental errors can no longer be considered as sufficient (2011). Along the same lines, Tim McNamara (1996) recommends

> to accept that the most appropriate aim of rater training is to make raters internally consistent so as to make statistical modelling of their characteristics possible, but beyond this to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some way. (p. 127)

Multi-faceted Rasch Analysis (MFRA) provides a helpful solution to model rater variation (whereby each type of effect is called a *facet*) and compensate for these sources of variations when estimating the ability of a test taker based on her or his performance. An MFRA transforms a test taker's raw scores; the *measure* of a candidate's ability that an MFRA provides "results from an automatic adjustment of the candidate's raw score to take account of what is known about the influence of these facets" (McNamara, 1996). Such an analysis can be conducted 1) to measure how all these factors (facets) interact with one another and how these interactions may determine the likelihood of particular test scores, 2) to control and compensate for them (McNamara, 1996), and 3) to avoid drawing erroneous conclusions when evaluating the psychometric quality of ratings (Eckes, 2011). As Eckes states: "Erst eine Multifacetten-Rasch-Analyse kann letztlich Aufschluss darüber geben, wie verlässlich und aussagekräftig die von Beurteilern abgegebenen Einstufungen sind" (Eckes, 2011). In sum, rater variability must not be underestimated in practice and should therefore constitute an important focus in any performance assessment. Once measures are undertaken to compensate for such variability, a test may call for a need to set a standard. What this means and how this can be achieved, will be explored in the following subsection.

2.5.5. Setting Standards

Another essential component of performance testing is the scale against which the language performances are judged. This results in the assignment of a language learner to a proficiency level based on said scale. Competence models with proficiency levels provide the foundation for identifying learners' competences through inferences based on observed language

performances. While assessment criteria provide the basis for rating discrete task performances, the ratings as a whole provide the basis for standard-setting and defining cut scores. Generally, trained raters are responsible for marking the language productions generated through a performance test, and a standardising committee is responsible for (a-priori or a-posteriori) setting the standard. Such standards or cut scores serve the goal of categorising language performance into proficiency levels. With education being increasingly competence- and standard-oriented, competence models gain significance because they provide explanations of the structure, proficiency levels and development of competences. In (language) teacher education, competence models represent an implicit attempt to establish proficiency standards for didactic and pedagogical actions of teachers based on standard setting procedures informed by education theories (Hallet & Königs, 2010). Proficiency levels can be deducted in two ways. A-priori-methods include setting standards and defining cut scores based on theoretical models prior to collecting empirical data. In contrast, a-posteriori-methods encompass deriving standards based on existing empirical data, i.e. test scores. The latter generally involve standard-setting procedures to identify cut scores. Cizek and Bunch (2007) define the concept of a standard-setting method as a process that requires group decisions and involves making informed, evidence-based judgements. As Cizek and Bunch (2007) explain, "[t]hese judgements are summarised in some systematic way, typically with the aid of a mathematical model, to produce one or more cut scores" (p. 65). In language testing, standard setting is defined as "the process of establishing one or more cut scores on examinations […] [where] the cut scores divide the distribution of examinees' test performances into two or more categories" (ibid. p. 65). In other words, standard setting is a system of rules or procedures that includes interpreting a performance standard as an operational position on a scale (Cizek & Bunch, 2007). For example, if test performances are to be classified into the three categories *Basic*, *Proficient* and *Advanced*, "two cut scores are needed – one to define the border between Basic and Advanced, and another to define the border between Proficient and Advanced" (ibid. p. 4). It is imperative that the standard-setting procedure aligns with the test purpose; hence, the test purpose needs to be defined, articulated and recorded in detail prior to setting the standard (Cizek & Bunch, 2007). As standard setting procedures provide a structured and reasoned approach that results in allocating a number to discriminate between at least two levels of performance, they need to be systematic and provide procedural and internal validity evidence (Cizek & Bunch, 2007; ALTE, 2020). In the case of LSP/LAPP testing, for example, setting the standard includes qualified experts following systematically developed procedures, usually

to interpret whether the elicited test performance provides evidence that the test taker's L2 abilities are sufficient to meet the language requirements of the respective professional domain (Manias & McNamara, 2016). The aim of following such rigorous measures is to minimise any chance of randomness and bias and to increase the reliability of decisions related to setting cut scores (Alderson, Clapham & Wall, 1995; Kenyon & Römhild, 2014). Especially in contexts where LSP tests are of high stakes, act as gatekeepers and/or afford professional or life opportunities, setting defensible performance standards is crucial (Knoch & Macqueen, 2020). Despite the relevance of achieving high validity in the standard-setting process, what constitutes an acceptable passing standard is essentially a judgement that reflects the values of the stakeholders who take part in its definition (Knoch & Macqueen, 2020). Consequently, a set standard can by definition never be entirely empirically correct (Knoch & Macqueen, 2020). In order to ensure the defensibility of significant thresholds and minimise arbitrariness and randomness of results, it is therefore pivotal that standard-setting workshops follow a principled set of procedures (Knoch & Macqueen, 2020). As Kenyon and Römhild (2014) point out, the procedural rigour in standard setting itself constitutes an element of the validation of the exercise. There are multitudes of standard setting methods (cf. Cizek & Bunch, 2007, for a comprehensive overview). They have been broadly classified into two types, those that are *test-centered* and those that are *examinee-centered* (Jaeger, 1989). In *test-centered* methods, standardising committee members focus on test tasks or test items and evaluate how a *minimally competent* test-taker (i.e. the least able candidate) would perform on each of these items (Cizek & Bunch, 2007). *Examinee-centered* methods involve expert judgements of actual test performances of test takers (Cizek & Bunch, 2007). The analytic (or analytical) judgement method (Plake & Hambleton, 2001) and the Body of Works method (Kingston, Kahl, Sweeney, & Bay, 2001) are examinee-centered methods commonly used for constructed responses such as speaking and writing tests (Knoch & Macqueen, 2020). The selection of a suitable standard-setting method is of high importance and requires taking a number of key features into consideration. First, the method needs to align with the test purpose and the complexity of the knowledge, skills, and abilities the test assesses (Cizek & Bunch, 2007). Second, the test design needs to be considered, as not all methods are applicable or adaptable to all item formats. Third, the number of performance categories (cut scores) required need to be taken into account. A final consideration concerns the amount of resources available for setting the standard. It is further imperative to be aware that method effects may occur irrespective of the selected method (c.f. Kane, 1994) which may threaten the validity of the resulting cut-scores. Finally, the

defensibility and validity of any standard-setting method depend highly on aspects that occur in advance of the actual meeting, such as standardising committee selection or quality of rating scale and PLDs (Cizek & Bunch, 2007). In sum, ensuring that standard setting procedures – *and* rater trainings – are conducted with appropriate rigour contributes to the overall validity argument of a test and can warrant the entire testing machinery adhering to orderly processes.

2.6. Summary

In the present chapter, I outlined the central theoretical concepts that underlie L2 oral speech production, teacher language competence and oral teacher feedback skills in L2 education to provide the necessary foundation for understanding the overall research questions and research project. Because this dissertation is mainly concerned with the development and assessment of profession-related oral L2 competences of pre-service L2 teachers, I focused on communicative competence as conceptualised in the CEFR including an action-oriented approach to L2 development within the socio-constructivist theory of learning. This provided the groundwork for making considerations on the central role of mediation as an influential concept in teaching and thus in teacher language competence. Within this setting, I explained approaches to conceptualising the construct of teacher language competence including its most elaborate and pragmatic realisation in the PRLCP and PRLC-R. Considering the high impact factor of the PRLCP in Switzerland, I outlined the need for further research in this area, thereby reinforcing the importance of the present dissertation. By zooming in on the PRLCP, I identified AoA 3 with its approach to feedback skills to be of special significance to L2 teaching and thus to teacher language competence and thus localised the area of interest to this research. By discussing the theoretical considerations of feedback within this context, I designated the contemporary paradigm of feedback based on the socio-constructivist approach to learning as the central theoretical understanding of feedback used in this dissertation. Further elaborations on feedback literacy including teacher and student feedback literacy reinforced the important role language and orality in the feedback process and in developing teacher and student feedback literacy. Subsequently, I outlined how L2 skills can be assessed including presenting some central concepts of language testing. This provided the basis for outlining current practices and challenges of assessing oral language competence in language for specific purposes and language performance contexts – challenges that this dissertation study seeks to address. Finally, I outlined ways of scoring teacher language performance including the challenges of human ratings. Based on these theoretical elaborations, I will now proceed to

reviewing the literature related to the development of student teachers' teacher language competence and feedback skills, and developing feedback literacy and L2 oral skills through feedback to then deduct the detailed research questions for this dissertation study.

3 Literature Review

1.1. Introduction

This chapter consolidates the research desideratum approximated in the above chapters (see chapter 1.3), contextualises the research aims and establishes the precise research questions by synthesising the current research findings on oral teacher language competence and identifying gaps in the literature. Accordingly, it builds on the conceptualisations of communicative competence (Bachman & Palmer, 1996; Council of Europe, 2001; Hymes, 1972), the construct of teacher language competence (Elder & Kim, 2014; Freeman, 2017; Freeman et al., 2015; Kuster et al., 2014), and teacher and student feedback literacy (Carless, 2020a; Carless & Boud, 2018; Carless & Winstone, 2020; Chong, 2021; Sutton, 2012). The literature review process was guided by the following question:

> What does the existing research reveal about pre-service teachers' ways of developing their L2 oral (teacher) language competence with particular reference to providing feedback or developing teacher and/or student feedback literacy?

In order to explore this question, I searched academic databases (e.g., EBSCOHost) to pull empirical research articles related to L2 oral language development in L2 teacher education and higher education. Next, I conducted a manual search and review of titles and abstracts in academic journals related to education, pedagogy, language assessment and applied linguistics (e.g., TESOL Quarterly, Language Teaching, Language Testing, Assessment and Evaluation in Higher Education, etc.). The following inclusion criteria guided the decision-making process for incorporating articles in the review:

- Research topic and outcome: communicative language teaching (CLT), methods for and studies on oral (teacher) language competence development, oral feedback

competence development, oral teacher and student feedback literacy, oral L2 development through (peer) feedback and rubrics

- Population / sample: pre- and in-service teachers, L2 learners in higher education
- Study design[15]: conceptual articles, intervention studies, experimental and quasi-experimental studies, action research, quantitative, qualitative, and mixed-methods approaches

Papers on L2 writing, recasts, corrective feedback, young learners, and language / English for academic purposes were excluded from the literature review because of the missing relevance and relation to the development of oral (teacher) language competence. In addition, studies of low quality were excluded. This encompassed studies that did not have a methodology or a results section. After selecting and studying the resources, I organised the findings from the literature according to treatment methodologies employed. This process resulted in the emergence of five analytical themes related to the research question. The overall themes indicate the methodological and thematic foci of the respective research studies and include:

1. oral L2 development through CLT,
2. oral L2 development through raising awareness (RA),
3. oral L2 development through (peer) feedback ((P)FB),
4. oral L2 development through multi-stage assessments, videos and reflection (MS/V/R), and
5. oral teacher language competence development.

Note here that the themes do not exclude one another but may overlap in multiple instances.

1.2. Overview

Before reporting the findings in relation to analytical themes 1-4, a first overview of the empirical studies selected for the literature review is presented in Table 3. Studies concerning

[15] Like in most fields of inquiry, empirical research in L2 education distinguishes between qualitative and quantitative approaches, thus assuming that qualitative and quantitative research are two mutually exclusive, diametrically opposed approaches. Critics claim that this view reduces the wide spectrum of possible research designs. Instead, they argue for conceptualising empirical research on a continuum (Grotjahn, 1987). Depending on the research interest and study design, empirical research designs may thus simultaneously employ a variety of methods (Caspari et al., 2016). Therefore, and for reasons of achieving a comprehensive literature review, I decided to include relevant empirical research from the entire continuum (cf. Schramm, 2016).

theme *5: oral teacher language competence development* are excluded from the below table because of their conceptually and structurally different nature. Instead, they will be described further below.

Author	*n*	Partici-pants	Theme	Design / Instruments	Dependent variable (DV)	Independent variable (IV)	Results
Salem Al-Yaseen (2020)	40	Pre-service EFL teachers	CLT	6-week pre-post-test quasi-experimental	English speaking skills	Cooperative learning task (jigsaw task)	E group improved speaking skills; positive impact on E's participation & enthusiasm
Köroğlu & Çakır (2017)	48	Pre-service EFL teachers	CLT	8-week pre-post quasi-experimental	English speaking skills	Flipped classroom	E improved speaking performance
Abdullah et al. (2019)	27	UG EFL students	CLT	1semester pre-post quasi-experimental	English oral performance	Flipped classroom	Overall improved speaking performances measured by IPAF test
Faez & Karas (2019)	69	Pre-service EFL teachers	CLT	1-year pre-post-exploratory, descriptive case study	English proficiency	1-year study abroad MA TESOL programme	Self-assessed proficiency increased overall by half a CEFR level across all scales
Gartmeier et al. (2015)	168	L1 medical students & L1 pre-service teachers	CLT	300-min. experimental, randomised controlled trial, 3 treatment groups, 1 wait-list control group, post-test	Professional L1 communi-cation skills	3 conditions of professional commu-nication module: (a) e-learning with video analysis of professional conversations, (b) role-play & (video) feedback, (c) combination	Condition (c) (combined) was more effective than condition (a) video-based e-learning and condition (b) role-play alone, condition (a) e-learning was more effective than condition (b) role-play
Muñoz Julio & Ramírez Contreras (2018)	35	Pre-service EFL teachers	RA	6-week action research, non-participant observation	English speaking skills	TCS	Improved speaking skills
Chan (2017)	37	UG Business students, NNSE	RA	1-semester action research / research-informed teaching	L2 spoken business English skills	Transcripts from authentic Business conversations to raise awareness of LSP spoken discourse features	Inconclusive evidence, some heightened awareness politeness strategies. Participants found transcript learning useful
Kissau et al. (2019)	15	Pre-service L2 Spanish teachers	RA	15-week action research, Post-test mock OPI, student re-flection, post-intervention interviews	L2 oral communication skills	Proficiency-based, interdepartmental online communications course	8 of the 15 participants' OPIc scores improved by at least one proficiency level, self-reported positive impact on Spanish oral proficiency
Gómez Sará (2016)	14	In-service teachers, beginner L2 English learners	(P)FB	11-session action research	development of the spontaneous interactive L2 speaking skills	PFB and corpus-based learning, Video-recorded speaking performances, feedback checklist, journal entries	Improved willingness to improve, use of compensatory strategies, construction of personalised corpus; underassessment & dependency on corpus
Rodriguez-Gonzalez & Castañeda (2018)	17	Inter-mediate L2 Spanish learners	(P)FB	14-week pre-post quasi-experimental study	Oral L2 Spanish skills	Trained PFB, video-recorded speaking performance, holistic rubric scoring & questionnaire	No improvement on speaking skills, self-reported positive perceptions of feedback for L2 development, higher degree of self-confidence & self-efficacy
Kırkgöz (2011)	28	1st-year pre-service English teachers	(P)FB	1-semester mixed-methods, pre-test needs assessment, post-test video-recorded speaking tasks	English speaking skills	Blended learning & peer feedback	Improved oral communication skills & positive perception of blended learning approach
De Grez et al. (2009)	57	1st year L1 Business Admi-nistration students	(P)FB	Pre-post quasi-experimental study, rated oral presentations & questionnaire (duration NA)	L1 oral presentation skills	3 conditions: PFB, expert feedback and self-observation	No significant impact of feedback mode on students' development of oral presentation skills, significant increase of overall oral presentation skills
Murillo-Zamorano & Montanero (2018)	32	L1 economics & business students	MS/V/R	2-3-week pre-post-follow-up quasi-experimental study	Quality of content & expressiveness of L1 oral presentation skills	Peer & teacher feedback, analysis of oral presentations by expert according to analytical rubric	PFB group improved significantly & more than teacher feedback group, follow-up test: results not maintained
Hung & Huang (2015)	NA	EFL students	MS/V/R	18-week pilot study	English oral presentation skills	Video blogging	Improved oral presentation performance

Cabrera-Solano (2020)	42	UG EFL students	MS/V/R	5-month study, pre-& post questionnaire	English speaking skills	E-portfolios with uploaded videos of speech productions, rubric-assessed speaking performances	Reported suitability of e-portfolios & conducive to learning and feedback; significant improvement of pronunciation & fluency
Lao-Un & Khampusaen (2018)	44	EFL UG nursing students	MS/V/R	NA	English speaking skills	E-portfolios including video-recorded speaking tasks and video-self-reflections, rubric-assessed speaking performances	Improved speaking ability, learner autonomy & media literacy skills, positive attitude towards intervention
Yeh et al. (2019)	45	EFL students	MS/V/R	Pre-post exploratory study (duration NA)	English speaking performance	Online peer feedback through blogs based on analytical assessment rubric	Improved scored speaking performances
Kennedy & Lees (2016)	19	L1 UG early childhood pre-service teachers	MS/V/R	1-semester study	Development of positive L1 adult–child interaction with infants and toddlers	Field-based placement with video-based performance feedback, weekly assessment of classroom performance using CLASS	Improvement of interactions & developmentally appropriate teaching behaviours, reflection & practice
Bower et al. (2011)	24	L1 pre-service teachers	MS/V/R	NA	L1 communica-tion competence, ability to interpret & analyse com-munication & effective presentation skills	Video reflection, online-questionnaire	Improvement in self-perceived presentation capabilities, improved understanding of communication concepts, reduction in communication anxiety, increase in confidence
Castañeda & Rodríguez-González (2011)	9	L2 university students	MS/V/R	1-semester study	L2 oral language performance	RSE against holistic rubric, self-evaluation and training intervention	Perceived increase of speaking skills & awareness

Table 3 : Overview summary literature review

Key:

CLASS	Classroom Assessment Scoring System	NA	Not available
CLT	Communicative language teaching	NNSE	Non-native speakers of English
DV	Dependent variable	(P)FB	(Peer) feedback
E	Experimental group	RA	Raising awareness
EFL	English as a foreign language	RSE	Retrospective self-evaluation method
IV	Independent variable	TCS	Transactional communication strategies
LSP	Language for specific purposes	UG	Undergraduate
MS/V/R	Multi-stage assessments / videos / reflection		

1.3. Findings

This section builds on the above overview and presents the findings of the literature review according to the five relevant analytical themes that emerged throughout the review process. The description of the studies is followed by a critical summarising discussion.

Communicative Language Teaching

The first analytic theme that emerged from the review concerns communicative language teaching approaches to fostering L2 speaking skills. One recent study within the CLT school of thought was conducted by Salem Al-Yaseen (2020), who investigated the effects of cooperative learning jigsaw tasks on 40 female Kuwaiti pre-service English teachers' L2 speaking skills, assessed according to the criteria fluency, accuracy, use of vocabulary, and pronunciation. A jigsaw-task usually involves group work including information gaps that need to be bridged. The quasi-experimental study included a pre- and post-test, which encompassed (more or less) profession-related tasks, such as presenting a teaching technique or an educational game.

Results show that the experimental group significantly outperformed the control group in the post-test in all relevant assessment criteria. The experimental group also displayed a change in attitude towards cooperative learning, showing that the treatment had a positive impact on students' self-perceived participation and enthusiasm. Other positive effects on language learners' speaking skills were observed in two studies that adopted a flipped classroom approach to develop language learners' speaking performance. For example, Köroğlu and Çakır (2017) implemented a flipped-classroom intervention designed to foster more classroom participation by implementing collaborative speaking activities and student-centered, active learning. The quasi-experimental, pre-post experimental-control-group design revealed that the experimental group significantly improved their speaking performance in the categories fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. Similarly, Abdullah et al. (2019) conducted a quasi-experimental research study with 27 undergraduate EFL students to investigate the effects of implementing a flipped-classroom approach on the learners' spoken L2 language performance. An adapted version of the Individual Presentation Assessment Form (IPAF) oral proficiency test was used before and after the treatment as a pre- and post-test. Three independent raters judged the participants' performances against the criteria self-confidence and assurance, body language, accuracy, fluency, and organisation. In addition, observations and focus group interviews were conducted to capture students' perceptions of the flipped-classroom model as well as of their self-perceived oral language performance development. Similar to Salem Al-Yaseen's (2020) findings, the results indicate that the participants improved their speaking performances significantly over the treatment period. Both Salem Al-Yaseen (2020) and Abdullah et al. (2019) conclude that student-centered, flipped-classroom instruction can be conducive to L2 oral skills development, and a combined consideration of these studies strengthens the evidence of favourable effects of such treatments. A slightly different, less "guided" approach to CLT was investigated in an exploratory, descriptive case study by Faez and Karas (2019). They investigated the perceptions of 69 non-native English teachers on their experiences in a one-year study abroad MA TESOL programme in Canada including its potential for enhancing their language proficiency (Faez & Karas, 2019). Pre- and post-stay self-assessments based on the global CEFR L2 proficiency scales (Council of Europe, 2001) as well as a post-stay reflective assignment reveal that the participants specifically benefitted from being immersed in English-medium instruction, learning about English-teaching- and language acquisition theories including *English for teaching*. Overall, their perceived language proficiency improved by half a CEFR level. The

participants also reported higher awareness of their own L2 skills and greater theoretical knowledge of language acquisition processes. The results suggest that these participants specifically benefited from language training that emphasised *English for teaching* rather than English for general purposes (Freeman, 2015, 2017). A quite different approach was taken by Gartmeier et al. (2015) who researched the relative effectiveness of 3 versions of a professional communication skills module on physician–patient and teacher–parent L1 dialogues. Both the assessment of conversation competence and the training programme were designed on the basis of a professional conversational competence framework established by Gartmeier et al. (2011). The researchers investigated the effectiveness of (a) e-learning including video-analysis of professional conversations, (b) role-play and (video) feedback, and (c) the combination of (a) and (b). To measure the 168 participants' L1 communication skills, trained raters coded the video-recorded role-plays of participants simulating an expert (teacher/medical professional) and trained actors simulating a layperson (e.g., parent/patient). The results showed that condition (c) (combined) was more effective than condition (a) video-based e-learning and condition (b) role-play alone. In addition, condition (a) was more effective than condition (b). The overall module (especially condition (c)), was significantly more effective for pre-service teachers than for medical students. Like with the previously outlines studies, the present findings speak in favour of the promising nature of CLT. Gartmeier et al. (2015) also report indices of favourable effects of observational learning through video-based e-learning on developing oral communication skills. While I decided to allocate this study to the CLT theme, mainly because of its role-play features, it may as well have been allocated to "raising awareness" in the next thematic strand, which I will now proceed to outline.

Raising Awareness

A range of action research projects investigated treatments that focus on raising learner awareness and thereby promoting oral L2 development. One such project was conducted by Muñoz Julio and Ramírez Contreras (2018) to investigate the effects of transactional communication strategies (TCS) on 35 pre-service teachers' speaking skills. The learning arrangement included a diagnostic stage and 6 subsequent workshops that were designed to teach, practise, reflect on and self-assess their use of transactional communication skills (*chunks*). After each workshop, the participants' language performance was assessed according

to a speaking assessment rubric adapted from the READI Oral proficiency scheme[16] (Finch & Sampson, 2004), including vocabulary and grammar, fluency, attitude, pronunciation, interaction, and transactional strategy use. The participants revealed significant progress in relation to vocabulary, grammar, fluency, attitude, pronunciation, and interaction.

Another approach within this thematic strand was taken in a higher education LSP business communication context. In an attempt to strengthen the link between research and pedagogy in Business English teaching, Chan (2017) took a research-informed teaching approach to foster L2 spoken business English skills among undergraduate Business students at a Hong Kong university. The researcher used transcripts from authentic Business conversations to find out whether these materials aided raising students' awareness of LSP spoken discourse features. The transcripts of authentic workplace talk were extracted from books written by academic researchers, and framed with in-class discussion questions and supporting tasks such as role-plays. The analysis of the participants' video-recorded role-plays showed indices that they had learned some business-specific language and communication skills, although the evidence was not conclusive. Some participants also demonstrated a heightened awareness of some language strategies such as politeness markers. Even though most participants deemed learning with transcripts as useful and interesting, it became apparent that familiarising students with transcription conventions and providing contextual information is indispensable to ensure that learners can benefit from such a learning arrangement. Kissau et al. (2019) took a different approach to raising awareness for developing L2 oral proficiency. In an experimental pre-post intervention design, they researched the possible impact of a proficiency-based, interdepartmental online course on 15 pre-service L2 Spanish teachers' oral communication skills. The course was collaboratively designed by faculty staff from the College of Education and the Department of Languages and Culture Studies of a US university in alignment with the ACTFL standards to foster the L2 speaking skills required to successfully complete the Oral Proficiency Interview (OPI). The course encompassed 15 weeks with a series of targeted teaching-to-the-test tasks to familiarise the students with the ACTFL Proficiency Guidelines (ACTFL, 2012) and the OPI. In addition, students completed a range of activities that involved practising two-way spontaneous communication, reviewing Spanish grammar forms, watching

[16] "The READI oral proficiency criteria are based on IATEFL criteria and the Canadian Language Benchmarks. They describe speaking ability in terms of: 1) **R**ange: vocabulary, grammar; 2) **E**ase of speech: fluency; 3) **A**ttitude: self-confidence, motivation, reduced anxiety/nervousness; 4) **D**elivery: volume, pronunciation, intonation, word-stress, speech-rhythm; 5) **I**nteraction: body language, communication strategies, social conversation skills." (cf. Finch & Sampson, 2004; emphasis mine. Retrieved from http://www.finchpark.com/books/u2u/cj/cj1htm/075_EMarkingCriteria.htm, accessed on 30.8.2021)

authentic TV shows, compiling lists of new vocabulary with descriptions to practice circumlocution, or recording and transcribing scheduled phone calls with peers. At the end of the course, the students completed a mock OPIc and reflected on their strengths and weaknesses in relation to the OPI requirements. Results show that 8 of the 15 participants' OPIc scores improved by at least one proficiency level – an increase that was statistically significant. Students who scored lower in the pre-test made the most progress, while students who scored highly did not improve as much. In additional post-intervention interviews, all of the 7 interview participants reported that the course – especially the assignments that involved watching TV shows or movies – had a positive impact on the development of their Spanish oral proficiency. The authors argue that these results provide compelling evidence that the interdisciplinary and collaborative design of a teaching-to-the-test course can contribute to higher L2 oral proficiency – at least as measured by the OPI. Whether this is transferrable TLU contexts remains unclear. This teaching-to-the-test methodology has similarities with another commonly implemented intervention to promote oral L2 development, namely learning arrangements that place a strong emphasis on rubrics. There is evidence of a variety of benefits of incorporating rubrics in a learning environment. For instance, research findings indicate that the use of rubrics in educational contexts has the potential to positively influence students' learning provided that they are trained in their application (Panadero & Jönsson, 2013). Several studies show that rubrics can support student learning through clarifying learning targets (Smit & Birri, 2014) and making assessment criteria transparent. They can also act as a supporting instrument for peer- and self-assessment during a learning process, improve students' self-efficacy, reduce anxiety, foster students' ability to self-regulate their learning (Panadero & Jonsson, 2013; Panadero et al. 2016; Saddler & Andrade2004) and develop assessment and (student) feedback literacy (Carless & Boud, 2018). Burch (1997) manifests that rubrics prove to be particularly beneficial in connection with portfolios as, due to the nature of the portfolio format, they allow progress-tracking through the analysis of work submitted at different points in time (Smit & Birri, 2014). Although researchers remain uncertain whether rubrics might have a direct effect on student achievement (Brookhart & Chen, 2015), the evidence of favourable effects of rubrics in an educational environment is promising. An example of investigating rubrics in L2 oral development can be found in Brian Radford's (2014) PhD thesis. He conducted an experimental pre-post-design study and found that L2 learners who were taught explicit language performance criteria increased their (human-rated and computer-scored) speaking proficiency over those learners who did not receive any teaching of criteria. Despite the highly positive

attribution to the use of rubrics and their learning potential, Bacchus et al. (2020) mention a word of caution by stating that rubrics that have been solely designed by instructors may not provide learners with sufficient clarity and transparency on test tasks. They conclude that this potential lack of clarity may be mitigated through using exemplars to support rubrics or through constructing rubrics in collaboration with learners themselves.

Feedback

The previous subsection on rubrics closely relates to the next analytic theme identified in the review. Indeed, rubrics and feedback are often researched in combination with one another, because feedback is often criteria-based. Thus, rubrics and feedback are often intertwined. In L2 development research, however, feedback and rubrics seem to yet be underinvestigated. This becomes apparent, for example, in the extensive literature review by Garbati and Mady (2015) to explore effective teaching and learning strategies to aid L2 oral development. Their aim was to distill their findings down to a selection of best-practice principles that can inform teaching practice. They identified that the following strategies were reported to be effective for developing L2 oral skills: "explicit teaching, scaffolding, providing authentic encounters, planned and spontaneous presentations, task planning, fluency activities, questioning, role-play, assessment and feedback" (ibid, p. 1764). While they found that most of the mentioned strategies quite commonly constituted the research object of empirical studies, feedback seemed to be present in only a few. Indeed, corrective feedback was the only form of feedback the researchers found to be reported on. Considering the prominence of studies on feedback in educational research and the widely recognised beneficial effects of peer feedback, it is striking that no other forms of feedback emerged as themes in Garbati and Mady's review. For instance, peer feedback is commonly considered a valuable approach to formative assessment (Falchikov, 2003, 2005; Falchikov & Goldfinch, 2000) and is said to increase active, collaborative and perceived learning with additional benefits related to communication, problem solving, leadership and self-management skills (Cheung-Blunden & Khan, 2018; Rafiq & Fullerton, 1996; Johnston & Miles, 2004; Panadero, Romero & Strijbos, 2013; Spatar et al. 2015). In addition, peer feedback is considered one of the key contributors to developing assessment literacy and, more specifically, student feedback literacy (Carless & Boud, 2018; Hoo et al., 2021; Mutch et al., 2018; Sutton, 2012). The assumption that students learn best when they serve as assessors, provide peer feedback, and in doing so apply rubrics, is widespread and supported by the findings of numerous empirical research studies (Dawson, 2017). So far, the benefits of peer feedback have preoccupied L1 and L2 education research,

especially when investigating its effects on learners' writing skills (Rodriguez-Gonzalez & Castañeda, 2018). Such benefits encompass increased confidence and reduced anxiety among language learners because the peer feedback process allows them to see and relate to their peers' work, strengths and weaknesses (Ferris, 2003). L2 learners are also said to develop an increased comprehension of themselves and their peers as writers (Berg, 1999; Hyland & Hyland, 2006; Paulus, 1999) and develop an enhanced awareness for their audience (Byrd, 1994; Scott, 1996). Through observing the peer feedback practices of their students, even teachers can benefit by furthering their own understanding of L2 writing instruction (Allwright, 2003). At the same time, learners' low L2 proficiency has been identified as an inhibitor, indicating that poor L2 skills may largely hinder successful peer feedback (Villamil & De Guerrero, 1996). Aside from insufficient language proficiency, further reasons for concern are low student feedback literacy, including learners' compromised ability to recognise valuable feedback by their peers (Leki 1990; Stanley 1992) and to revise their work accordingly (Connor & Asenavage, 1994; Liu & Sadler, 2003). Research also shows that L2 learners tend to provide surface-level and "rubber-stamp" advice that largely counteracts the potential benefits of peer feedback (Leki, 1990; Lockhart & Ng, 1993; Mendonca & Johnson, 1994; Min, 2016; Tsui & Ng, 2000). Finally, a large body of research continuously shows that students tend to prefer their teachers' feedback over their peers' and are consequently more likely to act on the former rather than the latter (Nelson & Carson, 1998). While all these insights stem from research related to peer feedback in L2 writing, research into the effects of peer feedback on learners' L2 speaking performance is still scarce. However, it is assumed that the benefits of peer feedback, especially in multi-stage assessments and process-based approaches to L2 writing, are transferrable to L2 speaking (Rodriguez-Gonzalez & Castañeda, 2018) – under the premise that the participants develop feedback literacy and receive relevant training (e.g., through teacher modelling or the use of exemplars) (Min, 2005, 2016).

This literature review reveals that research on feedback for L2 oral skills development has slightly increased since Garbati and Mady (2015)'s review. There are some lines of argumentation that indicate a tentative move towards investigating aspects grounded in the contemporary feedback paradigm with reference to L2 oral competence development. A promising approach to investigating the effects of peer feedback on oral L2 proficiency, for instance, was taken in a qualitative action-research study by Gómez Sará (2016). The study examined the effects of peer feedback, corpus-based instruction and task based learning activities on 14 in-service teachers' L2 speaking skills. The corpus provided a collection of

chunks in relation to language productions used in comforting situations that occur in educational contexts. Participants' video-recorded speaking performances, peer feedback checklists, peer feedback based on the "plus-minus-what's-next" approach (PMWN), and journal entries on the 11-session intervention were analysed employing the Grounded Theory method. The results reveal that peer feedback and corpus-based learning activities positively affected the participants' willingness to improve, fostered their use of compensatory strategies, and contributed to the construction of a personalised version of the corpus. While the treatment enabled the participants to reflect on their L2 oral proficiency based on the received peer feedback, to expand their vocabulary range, and to reduce language barriers, it also resulted in underassessment and participants' dependency on the corpus, which negatively affected their fluency and spontaneity in speaking. Overall, the study revealed no conclusive findings on the participants' L2 competence development despite some tangentially beneficial effects. Inconclusive and contrary results were found by Rodriguez-Gonzalez and Castañeda's (2018) study. In a multiple draft-based approach, the participants partook in peer feedback practices and role-plays with a potential interlocutor to practice L2 speaking. The researchers captured the participants' perceptions via pre- and post-intervention questionnaires, linguistically analysed the participants' video-recorded speaking tasks and the participants' feedback comments according to a holistic scoring rubric. The results indicate that the intervention did not improve the participants' speaking ability in terms of lexical variation and global accuracy. The researchers report that the frequency of self-corrections did not decrease, which they interpret as a lack of positive effect of peer feedback on L2 fluency. Since self-corrections are a natural and inherently central characteristic of speaking (cf. Luoma, 2009), it is questionable to what extent this measure should be used to make conclusive judgements about a learner's progress (or lack thereof) in L2 speaking. Despite the supposed lack of improvement, students' evaluations yielded positive results. The learners considered peer feedback as being of positive value to their L2 development. They also reported a higher degree of self-confidence and self-efficacy – results that correspond to studies on peer feedback conducted in relation to L2 writing. More positive effects on L2 speaking skill development are reported by Kırkgöz (2011) in a mixed-methods study investigating students' perceptions and effects of a video-enhanced blended-learning course that incorporated peer feedback in a task-based learning environment. The speaking course had been developed based on a needs assessment to identify the students' weaknesses, which simultaneously served as the pre-test. The course was comprised of three hours of task-based classroom instruction including a series of video-recorded speaking tasks

on topics relevant to the pre-service teachers' current and future needs. From an LSP/LAPP perspective, the chosen topics (money, education, language, fashion or tourism), however, mostly concerned *general* instead of profession-related language competence. In addition to the in-class speaking tasks, the research participants viewed and evaluated their peers' video-recorded oral performances. The analyses of pre-and post-test data, students' video-recordings, informal student interviews and a course evaluation show that the participants developed their oral communication skills and perceived the blended learning approach positively. What exactly these communication skill improvements entail remains unclear. The author concludes that filming and watching videos of one's own and peers' performances as well as reflecting on one's own videos positively impacted on students' ability to critically reflect on their task performance and skills – a finding that is consistent with peer feedback literature and feedback literacy conceptions as outlined above.

Besides studying peer feedback in L2 education, a prominent methodological approach within this thematic strand constitutes the investigation of the differential effects of various types of feedback on learners' oral L2 performances, often in combination with analytical scoring rubrics and multi-draft assignments. For example, Tseng and Yeh (2019) researched the differential effects of written and video-recorded, rubric-based peer feedback on 43 EFL students by capturing their perceptions of the feedback usefulness. The findings revealed that the students found written feedback useful when it came to improving their L2 accuracy. Students also reported that the video feedback was particularly insightful for developing their L2 intonation. Despite these positive reports, neither written nor video feedback had statistically significant effects on the learners' spoken L2 proficiency measured according to students' self-perceived fluency and pronunciation. Murillo-Zamoranoa and Montanero (2018) report a similarly positive influence of feedback on L2 speaking skills. In their study, the differential effects of peer and teacher feedback on the quality of Economics and Business students' oral presentation skills were investigated in a pre-post-follow-up (quasi)experimental design. Condition (1) encompassed peer assessment against a 6-scale analytical rubric for oral academic presentation skills. Condition (2) included immediate teacher feedback. An external communication expert analysed the 96 oral student presentations against the same 6-criteria rubric at three stages: the pre-test, post-test and follow-up-test. The rubric-peer-assessment and teacher-feedback profiles were compared and the results showed that the students who received peer feedback improved overall by more than those who received teacher feedback. Significant differences were also identified between the pre- and post-test scores for the rubric-peer-

feedback condition in all evaluation criteria, apart from *conclusion* and *visual support*. Students who had received teacher feedback only made significant improvements in the *expressiveness* dimension. The fact that the analytical rubric instead originates from verbal performance level descriptions may explain why, in contrast to previous research, peer feedback was more effective than teacher feedback. A follow-up test a month later however showed that the rubric-peer-feedback-group could not maintain their treatment-related improvements. Accordingly, a single session of rubric-based peer assessment and relatively short interventions (roughly 2-3 weeks) may not suffice for sustainable change. In contrast – and in stronger alignment with the existing literature – De Grez et al. (2009) found more supportive evidence of teacher feedback over peer feedback when investigating the impact of different feedback conditions on learners' L2 oral presentations. The treatment grounded on the assumption that oral presentation skills can be developed through observational learning and included participants delivering an initial oral presentation in the first phase. In addition, participants completed a four-step multimedia programme on oral presentation skills, which included, among others, instructional, observational and reflexive sequences as well as assessing peers' presentations. Phase 2 and 3 consisted of students delivering two more oral presentations, to each of which they received feedback one of three alternative types of feedback. Type (1) and (2) encompassed computer-generated feedback supposedly developed by either an expert or a peer, while type (3) involved the completion of a self-assessment with reference to their own presentations. Results from the expert ratings of the video-recorded presentations indicate that participants' overall presentation skills improved significantly from presentation 1 to presentation 3. Even though no significant differences between the feedback conditions could be observed, the findings suggest that those participants who received feedback from an expert made more progress in their oral presentation skills than those who received peer feedback. The least progress was observed among those who completed the self-assessment only. These findings are, however, not statistically significant. As these few studies show, feedback seems overall to have some beneficial effects on L2 oral skill development, regardless of the type. The research findings are however not always conclusive and results are often not statistically significant. The frequent lack of control group, small sample sizes or questionable language assessments depress the conclusiveness, stability and generalisability of such findings, as tentative as they may be.

Multi-Stage Assessments, Videos and Reflection

As already evident from the studies described above, research on learners' oral L2 development (including peer feedback) often includes the investigation of treatment effects of multi-stage

assessments. Commonly, such assignments are conducted via (video-)blogs, video-reflections on e-learning platforms. Indices for their promising nature with reference to L2 oral competence development become apparent, for example, in a pilot study conducted by Hung and Huang (2015). They investigated the effects of video-blogging on L2 learners' English oral presentation performance. The treatment lasted one semester and involved students uploading four presentation files and providing peer feedback. Findings showed that the students' oral presentation performance significantly improved in relation to the criteria pronunciation, intonation, projection, posture, introduction, conclusion, and purpose. Assignments like these are particularly prominent, even more so when they are implemented within portfolios. This is because portfolios constitute a prominent tool for examining the effectiveness of multi-stage assignments in relation to oral L2 development. Due to the largely heterogeneous use and the manifold purposes of portfolios in teacher education, a consensus-based definition of the tool is neither possible nor expedient (Gläser-Zikuda et al., 2020). However, the following conceptualisation by Paulson, Paulson and Meyer (1991) may serve as a working definition for the present context:

> A portfolio is a purposeful collection of student work that exhibits the student's efforts, progress, and achievements in one or more areas. The collection must include student participation in selecting contents, the criteria for selection, the criteria for judging merit, and evidence of student self-reflection. (p. 60)

Bearman et al. (2020) add that portfolios are repositories of student work that can be used to promote learning as well as to demonstrate progress. Recently, e-portfolios have received special attention. This is mainly due to their promising digital features as well as their allegedly enhanced benefits in contrast to the analogue counterparts (Bearman et al., 2020; Gläser-Zikuda et al., 2020). For example, research suggests that e-portfolios can foster the development of media competences (Ntuli et al., 2009). They may also benefit and contribute to the exchange and dialogue within professional learning communities (Gläser-Zikuda, 2015). Overall, portfolios can serve as an instrument of professionalisation that fosters increased reflexivity among pre-service teachers, supports their development of profession-related competences, and contributes to higher learner autonomy, self-regulation and responsibility (Gläser-Zikuda et al., 2020). In addition, compiling portfolios in teacher education enhances pre-service teachers' methodological competence and promotes the necessary skills to implement portfolios in their own teaching (Gläser-Zikuda et al., 2020). Portfolios can also be approached from a structural point of view, where they serve functions such as the compilation of (related) documents or

formative and summative assessment (Gläser-Zikuda et al., 2020). Finally, portfolios may contribute to the development of assessment literacy and feedback literacy (Carless, 2020a, 2020b; Carless & Boud, 2018; Carless et al., 2011; Chong, 2021). They are a suitable means of formative assessment that correspond with the changing needs of a rapidly changing digital world (Bearman et al., 2020). It is important to note, however, that there is a lack of substantial empirical evidence that supports these claims (Gläser-Zikuda et al., 2020). Indeed, most publications on this subject are of conceptual nature. Thus, the potential effects of portfolios in teacher education may at best be assumed. Nevertheless, portfolios constitute an instrument with high (innovative) potential due to their flexible and reflexive nature as well as their ability to establish a strong link to teaching practice (Gläser-Zikuda et al., 2020). One of the few existing studies on the effects of e-portfolios on learners' L2 oral competence was conducted by Cabrera-Solano (2020) with 42 undergraduate EFL students at an Ecuadorian university. Their mixed-methods investigation revealed that the participants' pronunciation and fluency improved over the 5-month treatment period. The intervention encompassed students curating an e-portfolio with uploaded videos of their recorded speech productions and speaking activities on various topics. The course conveners then assessed the speaking performances according to a speaking assessment rubric adapted from Cambridge English qualification tests and provided personalised feedback. In addition, a pre- and post-treatment questionnaire captured students' perceptions of the treatment and their self-assessed oral competence development. Almost all students agreed that e-portfolios were a suitable tool to practice speaking in the target language and they considered e-portfolios to be conducive to the feedback process. The ratings of the participants' video recordings revealed a significant improvement with reference to their pronunciation and fluency. In a similar design implemented by Lao-Un and Khampusaen (2018), 44 Thai EFL undergraduate nursing students video-recorded their speaking performances on four speaking tasks and video-self-reflections and uploaded them to their e-portfolio platform. They subsequently received feedback from the researchers on their speaking performances. The researchers evaluated the participants' video-recordings according to a speaking assessment rubric derived from CEFR descriptors and found that the e-portfolio format was effective in improving participants speaking ability, learner autonomy and media literacy skills. The researchers also coded the participants' self-reflection-videos and found that the participants generally had a positive attitude towards the intervention. While the positive results of both Cabrera-Solano (2020) and Lao-Un and Khampusaen (2018) are promising and strengthen the argument for the effectiveness of e-portfolios to promote learners' L2 oral

competence, the lack of control group does impede the stability and generalisability of the results. It can thus not be concluded that the reported improvement is attributed to the treatment alone, or whether natural language learning processes over the treatment period contributed to the speaking skill development independent of the treatment itself. This issue remains prevalent in Yeh et al.,'s (2019) explorative study that also found positive effects of online peer feedback on 45 EFL learners' L2 oral proficiency. The intervention consisted of a peer feedback training session including familiarising students with the evaluation criteria, and three subsequent cycles. Each cycle included the students' filming and uploading their speaking performances to an online blog, providing peer feedback, revising their performances based on the received feedback and re-uploading the optimised version. Finally, the participants completed a worksheet to reflect on their experiences of providing and receiving peer feedback via blogs. The researchers rated students' videos from their first and final blogging cycle against the same scoring rubric that the participants were introduced to in their peer feedback training session. The results suggest that the participants improved their scored speaking performances in relation to criteria describing the delivery of their speeches, but not in relation to vocabulary use and accuracy. Students who generally made more progress between cycle one and cycle three also demonstrated significant improvement in developing the content of their videos (introduction, supporting points and conclusion). While these results provide promising evidence into the effectiveness of trained peer feedback through online blogging on L2 learners' speaking performances, they need to be interpreted with caution. For example, it is to be debated to what extent the rating criteria and the rating process contributed to reliably capturing the learners' speaking proficiency despite the conducted rater training. The exploratory study however yields promising results to better understand and administer peer feedback processes in higher education that are intended to aid L2 learners' oral competence development.

Yet another video-based study was conducted by Kennedy and Lees (2016). Their intervention study was situated in an US undergraduate early childhood teacher education programme and aimed to investigate the impact of video-based performance feedback on L1 adult–child interactions with infants and toddlers. 19 pre-service early childhood teachers participated in a guided apprenticeship as part of a one-semester long learning module. The module involved (1) weekly evaluations on participants' developmentally appropriate teaching by using CLASS (Classroom Assessment Scoring System), (2) video-based narrative peer feedback on participants' uploaded video recordings of their classroom performance, and (3) universal (e.g., on-site learning experiences, explicit feedback on lesson plans), targeted (e.g., explicit feedback

on strengths and weaknesses) and intensive (e.g., individual improvement plans or conferencing for students with minimal progress) educator support. After the conclusion of the module, the pre-service teachers participated in interviews to reflect on their experience and overall progress with reference to their development as a teacher. The research findings demonstrate that the treatment positively affected the participants' interactions as well as their developmentally appropriate teaching behaviours. In addition, the peer feedback induced reflection positively influenced the participants' reflection and practice alike. Even though this research study is situated in the early childhood education context and does not focus on L2 teacher language competence per se, and again, lacks a control group to substantiate the findings, the findings are applicable to other contexts concerned with teacher-student interactions. One such example constitutes the study by Bower et al. (2011) who also reported beneficial effects of multi-stage assessments including video reflection on L2 communication competence. The treatment involved 24 pre-service teachers analysing communication scenarios, video-recording and uploading oral presentations in form of microteachings to an online blog, and making reflective observations on their communicative actions. Reflective peer feedback served to inform their next presentation attempt. 50 microteaching-video posts with self-reflections and 106 peer responses were uploaded online. Results from an online questionnaire show that the participants significantly improved their self-perceived presentation skills. The qualitative feedback and online commentaries also suggest that video reflection improved students' grasp of communication concepts. Furthermore, students mentioned decreased communication anxiety and increased confidence – a stable finding in the peer feedback literature. The authors interpret this result as "evidence for the interrelationship between the cognitive, behavioural and affective dimensions of communication" (ibid. p. 324). Because the questionnaire data only provide insight into students' self-perceptions, there is no evidence with regard to the impact of the reflective practice on the participants' actual performance and speaking competence. Finally, another study working with multi-draft assignments, video-recordings and feedback was conducted by Castañeda and Rodríguez--González (2011). In their small study with 9 university students, they employed a retrospective self-evaluation method (RSE) within a multi-draft assignment to investigate the effects of self-evaluation on L2 learners' oral language performance. Upon the completion of each draft, the participants completed an RSE, which asked them to reflect on each performance and rate themselves against a holistic rubric. After the third draft, the students underwent a training session to learn techniques for successful communicative performance. This training intervention included the research participants

evaluating video speech samples from students of a previous semester against the same evaluation rubric. After a final fourth video recording developed based on the insights from the training intervention, a post-implementation questionnaire was administered. The results show that the learners perceived increased L2 oral skills awareness and competence through the iterative submission and analysis of their own and their peers' speech drafts. Limitations like the small sample size, the lack of a control group and the lack of objective measurements of the participants' L2 speaking ability are similar to the studies cited above. Nevertheless, the study shows that using speech samples from "comparable L2 learners as a tool for students to develop attainable expectations and evaluate both successful and unsuccessful speaking strategies" (p. 495) can have favourable effects on L2 learners' language awareness. In conclusion, multi-stage assignments with some form of reflective practice, iterative video recording, assessment rubrics, peer feedback and training interventions seem effective in promoting learners' self- and other-assessed oral L2 competence.

Oral Teacher language competence

The review of the literature so far predominantly concerns L2 learners' general oral language proficiency development in a variety of contexts and through the application of a variety of methods. However, research employing LSP approaches and specifically investigating oral teacher language competence, i.e. spoken profession-related language skills of L2 teachers, is lacking. There is indeed almost no empirical research dedicated to this particular topic aside from conceptual and theoretical frameworks as presented in chapter 2.3. One of the few, very valuable contributions that touches upon the subject of teacher language proficiency comes from Laura Loder-Büchel (2014). In her PhD thesis, she researched the association of young learners' L2 reading, writing and listening performance with their teacher's measured language ability, feelings of improvement and contact with English outside the classroom. In addition, she researched whether the amount of classroom time dedicated to reading, writing, speaking and listening correlated with their students' L2 performance. Findings suggest that teacher language competence or teacher exposure to English do not determine learner performance in the first two years of English instruction. In fact, advanced speaking and grammar competence negatively correlated with learner performance. Instead, teachers' self-perceived continuous speaking skill development and time spent on specific language skill combinations in the classroom positively correlated with learner performance. Loder-Büchel (2014) summarises the findings of her study with a clear call for action:

> It has been all too easy up to now to attribute quality English language teaching to teachers' language proficiency. In the future, stakeholders in the English language education of young learners need to change their focus away from subject-matter knowledge onto pedagogical knowledge. In this specific case, we need to focus less on the "what" of the amount of teacher language knowledge and rather more on the "how" of teaching and learning the language. If we want pedagogical knowledge to come to the forefront, then more rigorous qualifications related to teaching practices (as opposed to subject-matter knowledge) and much more research are necessary to clarify the dynamic interplay between teacher pedagogical knowledge and motivation to learn and its impact on learner performance. (p. 190)

Pedagogical knowledge (PK) in this context seems to constitute the missing link between a teachers' L2 proficiency and the ability to teach that language. At the same time, Kuster et al. (2014) claim that the PRLCP provide this missing link by fusing PK – among others – with general language competences and thereby precisely describing what teacher language competence is comprised of. Loder-Büchel's call for action corresponds with Chambless' (2012) call for more empirical research:

> To improve student learning of [foreign languages], best practices must be established through valid and reliable empirical research, particularly research on the relationship between teachers' [target language] proficiency and the classroom practices that facilitate student learning in terms of the development of their oral proficiency. (p. 1589)

Chambless (2012) bases this call on a synthesis of findings from SLA research that indicate that the quantity and range of L2 input impact on student learning. According to these findings, what promotes L2 learning is ample exposure to meaningful, i.e. comprehensible input (cf. Krashen, 1981), and substantial possibilities "to create meaning and solve linguistic problems in speaking and writing" (ibid. p. 142). Her argument thus underlines the importance of a teacher's oral proficiency in the target language and the significance it carries in teaching effectiveness *and* student success. Although a teachers' oral L2 proficiency is not the only determining factor in student learning, Chambless argues that it is generally deemed invaluable for effective teaching. If teachers' oral language proficiency is understood as including teacher language awareness (TLA), Chambless' argument is supported by claims such as those by Lindahl and Baecher (2016) with reference teacher educators. Their research study examined the degree to which explicit attention to language was observable in the feedback provided by teacher educators (practicum or placement supervisors) to novice TESOL pre-service teachers

within supervision cycles. They also explored how the supervisors' focus on language during feedback interactions may influence pre-service teachers' pedagogical decisions. In their understanding, TLA encompasses 1) teachers' target language proficiency, 2) their grammatical knowledge, and 3) their capability to plan lessons that are engaging and supportive of the learner (ibid. p. 29). The results show that, overall, the supervisors' feedback contained very little focus on language. Furthermore, the findings suggest that supervisors with extensive TESOL experience (and presumably high TLA) provided feedback to pre-service teachers' lessons differently than supervisors with lower TLA. What exactly these differences are is not explicitly stated. However, the researchers conclude that it is crucial for teacher education institutions to ensure that teacher educators and supervisors "place language awareness at the core of what is being learnt and taught" (p. 37). This American perspective on oral profession-related language competences places its focus very strongly on teach*er* effectiveness rather than teach*ing* effectiveness. Also, in contrast to perspectives such as those by Loder-Büchel (2014), Freeman et al. (2009), Freeman (2017) or Elder et al. (2014), there seems to be a widespread lack of awareness that general language competences are likely to differ from profession-related language competences. Indeed, the emphasis lies almost exclusively on the former. Even when Chambless (2012) calls for striving for a better understanding between teachers' language competences and student success and thus asks for a better connection to the real-life classroom, the issue seems to remain obscured.

This connection was targeted in a Swiss "in-the-wild"[17] study that did take an LSP bottom-up approach to classroom language and teacher language competence. Maya Loeliger (2015) employed a corpus-linguistics approach to investigate the profession-related functional chunks as they appear in L2 primary school teachers' language productions in the L2 German (DaF) classroom. Loeliger filmed and analysed L2 German lessons in the Swiss canton of Fribourg and administered a questionnaire survey. Once the video material was transcribed and coded (double blind), Loeliger built a corpus where she compiled and sorted the language occurrences in question. Finally, she created wordlists of common occurrences and mapped the findings onto the PRLCP. In this final step, it proved possible to project the developed codes and sub codes onto the PRLCP. While all sub codes could be mapped onto at least one descriptor of the PRLCP, not all PRLCP descriptors could be allocated with a sub code from the textual data. These findings lend support to some systemic relevance and validity of the PRCLP as well as the concept of teacher language competence in general. That this LSP approach to L2 teaching

[17] "In the wild" here refers to the real-life, authentic context of the L2 classroom.

reaches beyond the target level of primary and secondary education is apparent in the following studies related to English-medium-instruction (EMI) settings at universities. It is important to note that within the EMI setting the L2 is not primarily the object of instruction, but the *medium* in which a given subject is taught. Similar to profession-related language competences of L2 teachers at primary and secondary school level, research has shown that general language proficiency provides an insufficient basis for assessing lecturers' specific English competences for their suitability to lecture in an EMI setting. For example, Klaassen (2001) identified that effective language behaviour neither affected student achievement nor predicted lecture clarity. It did also not correlate with teaching effectiveness or student learning. The only impact high lecturer language proficiency had was on students' *perceptions* of understanding. Similarly, Björkman (2011) reported that effective language use in EMI does not depend on high language proficiency according to their research findings. These findings are in line with what Laura Loder-Büchel (2014) identified, albeit in a different educational context. Further findings by Pilkinton-Pihko (2013) support the CEFR-CV's orientation away from the native-speaker ideal, reporting that in intercultural EMI settings, comprehensibility was more important than native-like language proficiency. Furthermore, Studer (2015) found that a "lecturer's ability to negotiate communicative-didactic rather than linguistic competence" largely determines students' EMI experience. To further investigate these findings, Studer et al. (2018) developed an observation protocol on language-related teaching competences with the purpose of evaluating language-related teaching performance. The ultimate goal of the protocol was to determine lecturers' suitability to teach successfully in English-taught programmes (ETP). The project included the development of the descriptors in relation to EMI-lecturers speaking skills and testing them in a Swiss BSc Business Administration course. The developed dimensions included formal linguistic descriptors directly derived from the CEFR as well as indigenous criteria such as communicative-didactic competence. For testing them, 6 expert raters applied the criteria in an experimental case study while observing 10 courses in 8 modules at 90 minutes each. Findings show that while indigenous criteria were commonly rated as relevant for assessing lecturers' profession-related English oral competence, they were also considered difficult to judge because they seemed too broad for assessment. After analysing the results, the team finalised the list of descriptors by distilling the original 24 descriptors down to five. The final descriptors reflect that ETP experts connect language-related quality EMI to "1. phonological control in L2, i.e. little accent, hearer-oriented speech rate and lively intonation (general language competence); 2. student comprehension (dialogic competence […]); 3.

explicit content structure (communicative-didactic competence […]); 4. L2-consolidation activities (language-didactic competence […]); [and] 5. opportunities for L2 use in classroom (language-didactic competence)" (ibid. p. 45). Interesting to note is the emphasis on "little accent" within "phonological control", which implies an (unnecessary, if not inappropriate) return to the orientation on a native-speaker ideal. Studer concludes that the developed descriptors constitute a novel approach to conceptualising the L2 as an object of learning in EMI-lecturers' language performance, and thereby highlights the convener's role as a language and communication facilitator. The results of this study were further researched in a compelling approach taken by Curtis Gautschi (2018). Based on the notion that students are key stakeholders and important contributors to the design and validation of trained-rater assessment tools, Gautschi administered student questionnaires with the aim to validate the trained-rater assessment tool specifically with respect to their linguistic, communicative and didactic competences. In order to do so, the 67 student questionnaire responses were compared with the trained-rater evaluations from Studer's (2018) study to identify items related to students' perception of the quality of EMI-lectures. Generally, the findings show that the overall relationship between student and rater judgements was uneven. It particularly stands out that the agreement between student and expert raters was higher on formal linguistic, competence-related criteria and lower on indigenous criteria such as communitive and didactic items. Overall, while these insights lend support to communicative and didactic competences being important for successful EMI teaching, they also show that individual items such as those of the communicative and didactic scales lack evidence. Thus, even though all five dimensions contributed to student perceptions of quality, further modifications to the descriptors are necessary as part of the iterative process of rubric design.

1.4. Summary, Gap and Study Rationale

The above review of the literature reveals that most of the studies conducted with reference to L2 development concern L2 learners who complete a language course to acquire general L2 proficiency in higher education or in professional development. In contrast, only few studies investigate the L2 oral competence development of pre-service or in-service language teachers, let alone the development of teacher language competence. In most cases, such studies concern learners' prepared speeches and oral presentation or communication skills measured against formal linguistic criteria such as vocabulary, fluency, grammar, and "correct" pronunciation. Occasionally, conversation-making criteria or criteria of oral presentations are used for

assessment. However, none of those criteria demonstrates a close proximity to the actual demands of the L2 classroom. Instead, they seem almost entirely removed from the teacher language competence construct itself. Investigations of oral teacher language competence measured against indigenous criteria – and by means of LSP test tasks specific to the demands of the L2 classroom and the TLU domain – are to the best of my knowledge almost nonexistent. This indicates that the traditional assumption of high general language proficiency predicting effective teaching still builds the foundation of most empirical studies. This also demonstrates that there is a need for further research that adopts an LSP and action-oriented approach.

A prevalent finding of the above review constitutes the growing number of studies that incorporate blended learning approaches and video-based interventions. Treatments in these studies often include some form of multi-draft assignment paired with reflective components, rubrics and different types of feedback. For example, several studies show some tendencies of favourable effects of formative multi-stage assessments on students' L2 speaking skills that include iterative cycles of video recording students' own speaking performances (e.g., microteachings, oral presentations, simulations of conversations or role-plays), providing peer feedback and engaging in reflective practice related to the received (peer) feedback. Some studies focus more on the differential effects of different types of feedback (e.g., expert, peer, self or computer-generated feedback) while others focus more on the effects of pedagogical interventions and methods (e.g., flipped classroom, CLT, student-centred collaborative tasks, task based learning, etc.). Yet others focus on the effects of treatments that are designed to raise awareness, such as teaching-to-the-test, learning chunks, studying "authentic" or "genuine" materials such as exemplars, the work of peers, transcripts and TV-shows, or working with corpora, rubrics, and SLA as well as language teaching and learning theories. However, most of the studies presented above present considerable limitations:

Studies	***n***	***t***	**Pre/ post/ follow-up**	**Expert scoring**	**Self-assessment**	**Analytic / holistic**	**Question-naire**	**Inter-view**	**Obser-vation**	**Control group**
Salem Al-Yaseen (2020)	40	6 wks	Pre-post	x		a	x			x
Köroğlu & Çakır (2017)	48	8 wks	Pre-post	x		a				x
Abdullah et al. (2019)	27	1 term	Pre-post	x	x	a		x	x	
Faez & Karas (2019)	69	1 year	Pre-post		x	a				
Gartmeier et al. (2015)	168	300 mins	Post	x						x
Muñoz Julio & Ramírez Contreras (2018)	35	6 wks		x		a			x	
Chan (2017)	37	1 term	Post				x			
Kissau et al. (2019)	15	15 wks	Pre-post					x		
Gómez Sará (2016)	14	11 sessions	Post	x	x					
Rodriguez-Gonzalez & Castañeda (2018)	17	14 wks	Pre-post	x		h	x			
Kırkgöz (2011)	28	1 term	Pre-post	x			x			

De Grez et al. (2009)	57	NA	Pre-post	x		a	x		
Murillo-Zamoranoa & Montanero (2018)	32	2-3 wks	Pre-post-follow-up	x		a			x
Hung & Huang (2015)	NA	18 wks		x		a			
Cabrera-Solano (2020)	42	5 mths	Pre-post	x		a	x		
Lao-Un & Khampusaen (2018)	44	NA	Post	x	x	a			
Yeh et al. (2019)	45	NA	Pre-post	x	x	a			
Kennedy & Lees (2016)	19	1 term	Cont. (weekly)	x		a			
Bower et al. (2011)	24	NA	Post		x				
Castañeda & Rodríguez-González (2011)	9	1 term	Post		x	h			

Table 4 : Overview of the methodologies and tools employed

As Table 4 shows, most studies are of quasi-experimental, experimental or action-research nature that draw their conclusions based on small sample sizes. Most of them also draw inferences on L2 oral skills development based on learners' self-assessments rather than objective forms of language testing. If standardised language assessments are used to collect data, they are mostly weak performance tests (cf. McNamara, 1996). This means that the participants' productions are mostly assessed against criteria that measure general language competence. Furthermore, the treatment periods rarely exceed the length of a semester. Given the brevity of these periods, it is at times surprising how large the observed treatment effects are reported to be, which provides grounds for treating the results with caution. In almost no instances delayed post-tests are used to investigate the sustainability of the treatment effects. Thus, even if a short intervention caused significant improvement of participants' L2 oral language skills, it remains unclear whether these gains remained stable over time. Finally, most studies did not include control groups. A lack of control groups in empirical research studies means that only within-subject effects can be investigated, resulting in the fact that any observable effects can never with certainty be attributed to the respective treatment. This renders such studies less conclusive and weak. Nevertheless, some promising research findings indicate that students' oral L2 competences can be promoted through communicative teaching approaches that place a strong emphasis on learner autonomy and incorporate multi-stage assessments, peer feedback, rubrics and reflective tasks. So far, however, there is considerable a lack of stable evidence supporting these tendencies. In addition, to the best of my knowledge there have been no studies conducted so far that investigate L2 oral (teacher) language competence development with particular reference to the PRLCP as well as in relation to peer feedback based on the teacher and/or student feedback literacy framework.

Finally, and with reference to the PRLCP and the PRLC-R, teacher education programmes carry the responsibility to support pre-service teachers in enabling them to move towards a successful

professional career. The programmes do so by equipping them with the wide range of competences and tools that will allow them to facilitate learning among their prospective students as evidenced in the review above. The tools, frameworks or constructs employed in most studies are often developed by researchers, teaching and learning professionals, subject experts or – less commonly so – field experts (see chapters 1 and 2.3.3). Studies that investigate the effectiveness of such materials and frameworks often take place at the level of teacher education rather than the actual classroom at the target level. Examining whether these interventions are effective in terms of making the participants "better teachers", however, only paint a partial picture – especially when seeking to uncover their true impact in the actual target classroom. Considering that teacher education curricula aim to educate future teachers to teach effectively, it is imperative that the curriculum content is mapped to and relevant for real-world teaching and learning contexts. Involving the target group population in the development and evaluation of teaching and learning frameworks, resources and methods may contribute to achieving a more comprehensive and holistic product. Indeed, especially in the context of language assessment, the learners as the target group constitute a key stakeholder alongside other pivotal bodies in the education enterprise, such as the institution, policymakers, or representatives of the post-education workplace (Gautschi, 2018). Thus, their perspectives should be included in validation considerations, especially when the purpose of an assessment instrument is to measure the quality of teachers' practices (or language competences, for that matter). Considering that any "meaningful testing should reflect the target situation" (Pilkinton-Pihko, 2013, p. 3), it seems a logical consequence to attribute a central role to the student perspective in the development of teacher (L2) competence assessment instruments (Gautschi, 2018). Thus, adopting a complementary bottom-up approach – similar to Loeliger (2015) – carries the potential to contribute to the (ecological) validity and effectiveness of such tools. As both the PRLCP and PRLC-R were developed by teaching and learning experts and are therefore restricted to the expert- and practitioner-view, applying these resources in real-life teaching and learning contexts and gathering data on how the target group population perceives them can provide valuable insights into better understanding, further refining and validating the tools.

1.5. Research Questions

The findings from the literature indicate the high relevance of distinguished feedback skills to successful teaching practice, learner empowerment and learner achievement. They also show

the lack of means to assess oral teacher language competence in L2 teacher education (i.e. language-specific oral feedback skills) in a valid, reliable and objective way. Finally, tertiary students generally display limited ability to provide clear and effective feedback. The research evidence also points to promising effects of student-centered and communicative language learning, multi-stage assessments, video-based language learning, reflective practice, and peer feedback on learners' general oral L2 skills, which leads to the assumption that these methods are transferrable to an LSP context to foster oral teacher language competence. In addition, the impact of implementing the PRLCP and PRLC-R in L2 teacher education and language testing contexts is yet unknown. Thus, this dissertation seeks to address these aspects by investigating the following research questions subsumed in the following three themes:

Investigating the Implementation of the PRLCP and PRLC-R in L2 Teacher Education

Based on the findings from the literature, the implementation of the PRLCP and PRLC-R in L2 teacher education is examined with reference to their potential effects on fostering L2 teacher competence as follows:

Research Question #1:

> How do qualitative, language-specific aspects of pre-service English teachers' oral feedbacks in the target language English provided to lower secondary school students develop under the administration of a profession-related assessment rubric and systematic feedback training?

Hypothesis #1:

> Through the iterative and repeated application of the PRLC-R and peer feedback, the research participants' oral profession-related language competences improve as measured against the PRLC-R criteria.

Investigating the Usability and Functioning of the PRLC-R in Language-Testing

The lack of means to validly, reliably and objectively assess oral teacher language competence in L2 teacher education and the PRLC-R being a proposed tool to fill this void, the PRLC-R need to be examined with reference to their usability, suitability and functioning from a language-testing perspective. Thus, in addition to the above overarching question, further investigations are necessary, leading to the following three questions:

RQ #2.1:

Do the raters differ in the severity or leniency with which they rate the test takers' performances?

a) Does each rater maintain a uniform level of severity, or do particular raters score more harshly or leniently than expected?

RQ #2.2:

Do the raters maintain a uniform level of severity or leniency across criteria and across tasks?

a) Do ratings on one criterion follow a pattern that is markedly different from ratings on the others?

RQ #2.3:

Do the raters show evidence of differential rater functioning related to test takers' gender; that is, do they maintain a uniform level of severity or leniency across male and female test takers?

Investigating the Systemic Relevance of the PRLCP and PRLC-R in Secondary Schools

Just as the impact of the PRLCP and PRLC-R on teacher education and pre-service teachers' teacher L2 competence is yet unknown, so are their cascading effects on L2 teaching in secondary schools. At the same time, the instruments lack the scrutiny of a key stakeholder: the learners at the target level, i.e. field experts. Thus, research question #3 and its three sub-questions address the following:

Research Question #3:

How do lower secondary school students perceive and evaluate the linguistic quality and comprehensibility of pre-service English teachers' oral feedbacks in the target language English?

RQ #3.1:

How do lower secondary school students perceive and evaluate pre-service English teachers' English competence based on oral feedback performances in the target language English?

RQ #3.2:

What (language-specific) aspects of oral feedbacks in the target language English do lower secondary school students perceive as being crucial for ensuring student understanding?

RQ #3.3:

How do lower secondary school students' perceptions of pre-service English teachers' oral feedbacks in the target language English compare to those of trained experts in applied linguistics and English language teaching and learning?

Because research question #3 is highly explorative and I seek to adhere to the principle of openness in qualitative research, I refrain from formulating hypotheses to RQ #3. The above research questions #1, #2 and #3 including their respective sub-questions provide the framework for the current study. I aim to both gain insights into the application of the PRLCP and PRCL-R in the teaching and learning context of both teacher education and the target level classroom, and into the concept of teacher language competence on a broader level. This study constitutes an exploration of the affordances and limitations of the PRLCP and PRLC-R when applied in teacher education and comparable contexts, and an exploration of what these may mean with reference to L2 teacher education, the construct of teacher language competence and language testing. The subsequent chapters are dedicated to the description of the main- and sub-study including the methods for carrying out the study with reference to each of the research questions, the respective limitations, data analyses and results, discussion and implications.

4
Research Methodology Main-Study

This dissertation empirically investigates the implementation of the PRLCP and the PRLC-R in the Swiss L2 teacher education and language-testing context by conducting applied research and implementing an explorative, quasi-experimental, pre-post experimental control research design. The research study is embedded in a context-specific, applied-research environment that is of predominantly quantitative-descriptive nature (main-study, chapters 4, 5 and 6). In addition, it includes a supplement of qualitative and inductive research methodology (sub-study, chapters 7, 8 and 9). This present chapter describes the methodology of the main-study. First, the research design of the main-study is introduced through an overall contextualisation of the main-study (4.1). I then describe the research participants (4.2), research instruments including the pre- and post-test development (4.3), intervention design (4.4), and data processing and scoring procedures employed to answer RQ #1 (4.5). I conclude the chapter with a summary of the research methodology of the main-study and provide a transition to the subsequent presentation of the statistical data analyses and results.

4.1. Context

The research design of the main-study involves an intervention that applies the PRLCP and PRLC-R and a pre- and post-test to investigate their potential effects on pre-service English teachers' oral profession-related language skills on the example of oral feedback. It roots within a socio-constructivist theory of learning, which understands that development processes of complex skills occur through the negotiation and co-construction of meaning. To answer RQ #1, the investigation builds on the concepts of oral teacher language competence and teacher and student feedback literacy. It thus takes a student-centered approach with a focus on high learner autonomy, self-directed learning and the co-construction and negotiation of meaning. Pre-service teachers' oral profession-related L2 competences constitute the dependent variable (DV), and the PRLC-R constitutes the independent variable (IV). The study was conducted at

the St.Gallen University of Teacher Education (PHSG), home to the Center for Profession-Related Language Competences[18] that developed the PRLCP and the PRLC-R. The PRLCP and the PRLC-R are firmly embedded within the PHSG L2 teacher education curriculum. One such example is the Bachelor E-Portfolio (BA E-Portfolio), an already existing curricular component that implements the PRLCP to foster students' profession-related language competences over the course of one academic year. The multi-stage assignment is a practical equivalent to a bachelor's thesis and is introduced and facilitated in the mandatory 4th-semester course *Introduction to Linguistics*. It contains part A and part B as two separate tasks. The core objective of part B constitutes the development of students' oral, profession-related language competences in the target languages English and French by means of conducting, video recording, reflecting on and improving microteaching sequences in a learning-group environment. This process is paired with peer feedback and reflective practice. It is common for microteachings to be used as a tool for educator and peer feedback in teaching practicum modules (Joseph & Brennan, 2013; Ostrowski et al., 2012) and as context for peer feedback in e-portfolios (Joseph & Brennan, 2013). Such tasks usually involve pre-service teachers identifying learning goals for their microteachings, and their peers and educators providing feedback after viewing the (often video-recorded) microteaching (Kennedy & Lees, 2016). Accordingly, and in its original format, task B of the BA E-Portfolio contains three iterative steps that are repeated in three cycles:

1) In step one, each learning group (comprised of three to four students) composes a microteaching sequence of which each member has to perform a designated part. By devising the microteaching sequences, the students practise and display their oral, profession-related language competences in English or French. These microteaching sequences are audio- or video-recorded (i.e. multimedia sequence) and uploaded to SWITCHportfolio[19].

2) Within the respective learning group, each student provides peer feedback to another learning group member on her or his oral, profession-related language competences in step two.

[18] See https://www.phsg.ch/en/services/fachstellen/center-teachers-language-competences for more information (last accessed: 15.6.2021)

[19] SWITCHportfolio is a service offered by the SWITCH Foundation (SWITCH Information Technology Services; the Swiss national research and education network organisation). It is an e-portfolio system that provides a platform for students blogs and offers organisational, communication and reflection features to create and manage learning artifacts and progress monitoring. See https://portfolio.switch.ch/view/view.php?id=76056 for more information.

3) For step three, the students use the received feedback to individually reflect on their own oral, profession-related language competences and implement the suggestions in the subsequent microteaching sequence.

Throughout the entire e-portfolio process, each learning group receives guidance and supervision by a designated lecturer. The supervisor constitutes both the advisor and the assessor and provides the final mark on the assignment. The final e-portfolio part B product of each student contains three video- or audio-recorded microteaching sequences, three feedbacks on a peer's microteaching sequences, a transcript of one of her or his own sequences, and reflections of the complete development process. The BA E-Portfolio task B provides a suitable environment for the planned experimental-control intervention study to implement and investigate the use and potential learning effects of the PRLC-R on students' oral teacher language competences. Before describing the research instruments in detail, the next section provides an introduction to the research participants of the main-study.

4.2. Research Participants

The PHSG cohort 2017-2022 constituted the research sample for the main-study and included all students who chose to gain their teaching certification in at least one language subject. The group of 50 students (32% male, 68% female) consisted of two types: students with a focus on language and historical subjects who aspired to attain a Master of Arts (MA) teaching degree (phil I, n=40), and students with a mathematics and science focus who aspired to attain Master of Science (MSc) teaching degree including the teaching certification in one language subject (phil II, n=10). Because of their subject specialisation, completing the BA E-Portfolio is mandatory for all phil I students. The phil II students are exempted from the assignment and complete their e-portfolio equivalent in a science subject. Like the phil I students, however, they need to attend the course *Introduction to Linguistics* because of their choosing at least one language subject in their studies. All of the 50 students attended the course *Introduction to Linguistics* in the spring term of 2019. At the outset of the course, the participants were in their fourth semester of their secondary school teacher education studies, and in their sixth semester when they submitted the completed assignment in April 2020.

4.3. Research Instruments

This section describes the various research instruments in two main sections. First, I depict in more detail the PRLC-R as the core instrument of the dissertation. As the assessment rubric not only constitutes one of the most central components of the intervention design but also of the pre- and post-test, it also plays an integral part in the subsequent section that contains a detailed test description. There, I outline the individual test-development steps from the initial conceptual phase to the piloting stage to the final implementation. I also present attempts to overcome common issues of LSP tests such as their "limited ability to represent fully the demands of the target language use situation" (O'Hagan, Pill & Zhang, 2016). This includes a description of the measures taken to increase the degree of authenticity including the rationale for and development of video-vignettes.

4.3.1. PRLC-R Recap

The PRLC-R constitutes the heart of the present study and provides the basis of the intervention treatment as well as for the pre- and post-test. Developed based on the PRLCP, the instrument was designed for pre-service teachers, in-service teachers and teacher educators alike to facilitate the teaching, learning, and formative and summative assessment of profession-related language competences of L2 teachers (see chapter 2.5.4.3). The PRLC-R provides a coherent set of criteria, performance levels and performance level descriptors (PLDs, see also chapter 2.5.4; cf. Brookhart, 2013). The performance levels include a pre-entry level (*Vorstufe*), an entry-level (*Einstiegsniveau*), an intermediate-level (*Niveau en route*), and a professional-level (*Praxisniveau*). For devising the intervention, I extracted the scale *qualitative characteristics of speaking* of the overall PRLC-R. This scale contains the following assessment criteria: *task completion, vocabulary, accuracy, pronunciation, fluency, cohesion and coherence*, and *addressee-specificity* (see also chapter 4.2). Characteristic of these assessment criteria is, (with exception of the dimensions *task completion* and, to some degree, *addressee-specificity*), that they solely encompass formal linguistic aspects of language. Although the developers of the PRLC-R argue that the outlined linguistic assessment criteria are *profession-related* and thus represent *indigenous* criteria to allow for a profession-related lens on performance, they in actuality construe a "face resemblance", a mere impression of profession-relatedness. In a performance test that builds on the PRLC-R, task-elicited performances can thus only be interpreted in language-related terms. In other words, the focus of such a test can only lie on

the *language performance* itself rather than *task* performance (McNamara, 1996; Messick, 1994). This PLRC-R characteristic consequently only allows for the development of a *weak* performance test according to McNamara's (1996) classification (see chapter 2.5.2.2). This means that the purpose of the test instrument in this context thus can merely encompass "to elicit a language sample so that second language proficiency, and perhaps additionally qualities of the execution of the performance, may be assessed" (McNamara, 1996, p. 44). The PRLC-R are in line with Bachman and Palmer's (1996) recommendation to define and use criterion-referenced scales for performance tests because such scales enable drawing inferences on the language ability of test takers rather than only on the quality of their performance in relation to other learners (Bachman & Palmer, 1996). As the pre- and post-test of this study needs to enable the former as opposed to the latter, the PRLC-R are used accordingly. The subsequent section outlines the test development including the role and use of the PRLC-R as presented here.

4.3.2. Pre- and Post-Test

Pre-service English teachers' spoken language productions constitute the primary data to be collected in order to answer RQ #1. To access this spoken data by means of a pre- and post-test, an appropriate instrument needed to be implemented. With the research interest being on oral feedback skills, and due to there not being any known tests that enable collecting data on this *particular* construct from the PRLCP, an online-administered, criterion-referenced, near-authentic, competence-oriented performance test that elicits "constructed responses" in form of test-taker performances needed to be designed. Developing a second language performance test that is reliable, valid, objective, technically sound, practically useful, fair, and relevant for the test takers, involves following a rigorous method of a typically iterative nature (cf. for example ALTE, 2018; Bachman & Dambök, 2018; Knoch & Macqueen, 2020). The following sections outline the specific steps undertaken during the test development phase and provide a rationale for the decisions made throughout the process.

4.3.2.1. Test Development

The purpose of the LSP test that needed to be developed was to measure and assess the progress of pre-service teachers' oral feedback competences in the target language (TL) English. Thus, it needed to enable to determine the research participants' abilities to provide intelligible and precise feedback to lower secondary school L2 learners. The test also needed to be able to show whether the participants' respective abilities changed between t0 and t1 of the intervention.

Relevant language testing literature, guides and reference materials (ALTE, 2018, 2020; Bachman & Dambök, 2018; L. F. Bachman & A. S. Palmer, 1996; Douglas, 1997, 2000, 2010; Harsch, 2016; McNamara, 1996), and correspondence with language testing experts guided the test development and provided the necessary foundational knowledge from the initial conception to the blueprint to the piloting and finally to the final implementation. Informal consultations with language teaching experts and teacher educators served to gain a more profound understanding of the communicative demands of pre-service and in-service L2 language teachers in Switzerland (defining the TLU domain, cf. Jones, 1979). In addition, an online-LSP test that had been designed for a PHSG-internal research project to measure the spoken and written profession-related language competences of graduating students[20] served as the conceptual basis. Because this predecessor was also based on the PRLCP, some of the domain sampling and considerations of resources and constraints had already been conducted at a previous stage, which in turn informed the development of the pre- and post-test of the present study as well as the conception of the blueprint. In contrast to its predecessor, which measures both written *and* spoken language competences of all PRLCP AoA, this test solely serves the assessment of specific oral language skills taken from *AOA 3: Assessing, giving feedback and advising* (Kuster et al., 2014). The following 4 descriptors represent the test construct (see Appendix A for all AoA 3 descriptors and Appendix B for task specifications):

In the target language, the teacher is able to…

> 3.7 comment on the performance of a class.
>
> 3.8 conduct a dialogue that serves to assess a learner's ability.
>
> 3.9 give oral feedback on a learner's performance.
>
> 3.12 hold an advisory talk with learners with the aim of fostering their skills in a personalised manner.

These descriptors were extracted from AoA 3 because in terms of test practicality and feasibility they seemed the most suitable to be translated into test tasks. In addition, they cover areas of ability that are of particular relevance to (dialogic) feedback practice in the L2 classroom when approached from the socio-constructivist perspective as conceptualised in the idea of feedback literacy (see chapter 2.4.3).

[20] Consult https://www.phsg.ch/de/forschung/projekte/sprachstanderhebung-bei-studierenden-der-phsg for more information (accessed on 3.9.2021). As of the time this dissertation is submitted, there are no publications on this particular project and the developed LSP test.

Test Items

At the heart of performance assessment lies the development of test items that will elicit spoken language performances based on which the test takers' language ability in contexts outside the test can be inferred (Douglas, 2010). In alignment with the test purpose lay the decision that the present test was to be online-administered and criterion-referenced with *integrative* and *integrated* test tasks (see chapter 2.5.1). In order to design a near-authentic competence-oriented performance test based on the outlined test construct, the outlined 4 descriptors needed to be operationalised by turning the corresponding real-world tasks into test tasks. This needed to be done in a way so that the test items allow for reliable, objective and valid measurement of the language productions and for enabling reliable inferences to the underlying competence. Therefore, it was imperative that the test items constructively align with reliable and valid assessment criteria; here: the PLDs from the scale *qualitative characteristics of speaking* extracted from the PRLC-R (see Appendix C). In order to operationalise the test construct and draft test tasks, the following steps were undertaken:

1) consulting the test construct (4 descriptors AoA 3 of the PRLCP) and identifying prototypical sample scenarios that best represent each focalised area of competence. This selection process resulted in the decision to develop 7 test tasks;

2) selecting relevant criteria from the PRLC-R for assessing the language performances elicited through the prospective test tasks (i.e. the prototypical example scenarios). The criteria are *task completion, vocabulary, accuracy, pronunciation, fluency, cohesion and coherence*, and *addressee-specificity*;

3) drafting task specifications based on the sample scenarios and a template adapted from Bachman and Dambök (2018) (see Figure 13);

4) translating the identified sample scenarios into preliminary test tasks and composing scripts that reflect the each respective scenario (see below). Because the test was to be administered to French *and* English students, each task was developed in a French and English version. In this process, close attention needed to be paid to ensure comparability, feasibility and authenticity;

5) seeking feedback from experienced L2 teachers and specialists in didactics on the preliminary test tasks and scenario scripts to ensure that they closely represent authentic contexts, that the instructions are clear and that the tasks themselves are feasible, usable and understandable. This consultation served to collect what

Douglas (2010) refers to as *secondary data* on the first drafts of the test items and sample scenarios;

6) incorporating the feedback in the scripted classroom scenarios and the preliminary test tasks, and adapting the task specifications;

7) composing the final version of the respective scripts for the (near)authentic classroom scenarios to provide the foundation for selecting genuine test task stimuli and for creating video- and photo-vignettes. The completion of this step then provided the basis for the subsequent pre-pilot and pilot testing.

To meet the prerequisite of performance assessments, the respective TLU tasks require to be interrelated "in terms of the setting, the communicative goal to be achieved, and the participants" (Bachman & Palmer, 1996, p. 63). To achieve this for the present test, I constructed each test task within the following contextual setting:

> Welcome to this speaking assessment (Sprachstanderhebung). For the following tasks, imagine that you are substitute-teaching (stellvertreten) three English classes for a colleague who is currently away. You have only just started your substitution and are still getting to know your pupils. The scenarios you will see are all taking place in this context.

This description was to serve as a warm-up activity before the actual test. I phrased the warm-up task in the target language to provide the necessary context and help the test takers cognitively adjust to the specific language tasks and setting. The language of the task instructions were kept in German. The reason for this decision was to fulfil the main purpose of instructions, namely to ensure that the test takers could understand exactly the procedure of the test, test tasks and required responses (Bachman & Palmer, 1996). In other words, assessment instructions are not part of the test and should not jeopardise or inhibit the test taker's understanding (Bachman & Palmer, 1996). Additionally, I constructed the instructions to correspond to the nature of spoken language production (e.g., by avoiding instructions such as asking participants to speak in "complete sentences"). Bachman and Palmer (1996) recommend that instructions should correspond to the channel that is to be assessed. When assessing speaking competences, thus, the assessment instructions should be presented through input in the aural or audio-visual channel, i.e. through spoken instruction. For practical reasons, to avoid triggering biases on the test takers' side and to avoid misunderstandings caused by aural impairments, I decided to present the assessment instructions in writing. Finally, after the

completion of step 7 outlined above, the test development process proceeded to designating and creating the necessary stimuli for the respective test tasks.

Simuli

As performance assessment is characterised through the close relationship of the test tasks to the *real-world* (McNamara, 1996), the present test needed be of high authenticity, i.e., it needed to replicate the challenges and standards of performance that the test takers will typically encounter in their occupational real-life contexts (see chapter 2.5.1.2) (L. F. Bachman & A. S. Palmer, 1996; Caspari et al., 2016; Douglas, 2010). The incorporation of specific, near-authentic scenarios provides substantiated contextualisation and offers a close-to real-world communicative context. One way of creating near-authenticity in scenarios is through providing appropriate stimuli to elicit a specific reaction i.e. language production. Because of the affordances of video-based testing (see chapter 2.5.1.2), video- and photo-vignettes and genuine text-material were chosen to serve as real-world stimuli. Test takers needed to respond to those stimuli by recording an oral speech production – a prerequisite to enable any inferences from the observed test taker performance on their underlying competences (L. F. Bachman & A. S. Palmer, 1996; Caspari et al., 2016; Douglas, 2000, 2010).

Video- and Photo-vignettes

The stimuli chosen for each test task including all video-vignettes needed to match the predefined scenario content of each test item (see above). Campbell (1996) lists three major steps for creating vignettes: determining the issues and areas of concern, developing scenarios that are realistic and relevant, and testing the vignettes on groups similar to those who will be using them. Script writing involves a series of decisions that assist to ensure the development of an objective and valid performance test. For video-vignettes, such decisions include the definition of the classroom section to be captured, the length of each video and the number of videos required, the precise content, as well as the pupil actresses and actors and what they are to say and do (cf. Kaiser et al., 2015; see task specifications as illustrated in the development process outlined above). I decided that the individual video clips were not to exceed 30 seconds in order not to strain test takers' attention span. The filming of the video-vignettes was to take place with two groups of roughly 10-12 2nd-year secondary students in a lower-secondary school in Eastern Switzerland during two scheduled single lessons. The filming needed to be carefully planned, taking into account potential challenges of video-vignettes. Aside from the affordances of implementing video-vignettes in performance test tasks, Kaiser et al. (2015)

identify two main limitations that can compromise the validity of the test. They call the first of these two limitations the *twofold as-if*:

> Instead of experiencing the real situation, a certain video sequence is shown to the test persons as if it were real. Although it is obvious that there are differences to real classroom situations, it is difficult to name these differences completely. One of these differences [is that] the focus on a specific detail is already done in advance by the camera. So one important characteristics of expertise—namely distinguishing relevant from irrelevant aspects and focusing on the latter—[cannot] be measured in a completely satisfying way. (p. 384)

To minimise this problem, I used a wide camera angle while filming and ensured that the filmed actions took place in random areas in the classroom (thus, random locations in the video) to allow test takers to choose their own focus while watching the clip. The second *as-if* concerns test tasks that are referred to as *classroom-acting items*. Such items constitute a technique often applied in conventional vignette-based competence assessments. Traditionally, these test tasks confront the test takers with a classroom situation that includes a teacher-student interaction. The participants react to the stimulus by describing how they themselves would act *if they were* in the displayed situation. Such items diverge considerably from an authentic real-life classroom situation. As Kaiser et al. (2015) note: "By describing an action plan, another level of consciousness is involved in contrast to acting spontaneously and intuitively, the latter often done without being able to explicate one's own plans" (p. 384). To counteract this problem, video-vignette test tasks ideally prompt test takers to react *spontaneously* and *intuitively*. I mitigated this challenge by removing any teachers from the video and filming the scenario from the *first-person-perspective*. This should allow for a direct confrontation of the test takers with the classroom situation and enable a more spontaneous, intuitive reaction, moving from *act as if it was you* to *in this situation, it is you*. The following table summarises the problems associated with video-vignette-based test items and the solutions employed to mitigate them as appropriately as possible, *all things considered*:

Problems	**Implemented solutions**
Compromised authenticity, through e.g.,	a) Filming videos from a 1st-person perspective to simulate the perspective of a teacher in a classroom

a) Camera angle / perspective that forces test takers to look at whatever is in focus instead of choosing their own like they would in a real-world classroom b) Twofold-as-if	b) Filming videos with a broad camera angle to allow test takers to choose their focus on the classroom-situation more autonomously c) Assuring there is no teacher visible or audible in the video; the test takers should thus perceive themselves to be the teachers of the students depicted in the video d) Eliciting near-authentic responses through instructing test-takers to respond to the perceived stimuli through videos instead of instructing them to describe a situation ➔ However, authenticity is still compromised through the mere assessment condition: test takers complete the test while seated in front of a computer, and perceive information through videos and auditory stimuli through headphones
Difficulty to ensure comparability of test takers' receptive processes and perceptions as well as their individual processes of interpreting a situation	Providing precise and explicit task instructions that outline in detail what the test takers is to do
Lack of contextualisation / missing context (Kaiser et al., 2015)	Providing precise, extensive and clear descriptions of the context before test takers watch the video-vignettes

Table 5 : Problems of and solutions to video-vignette-based testing

After the initial planning phase, I video-recorded the scenarios in the classroom. To ensure efficient and successful filming within these time constraints, I piloted the handling of the filming equipment, the feasibility of the scenarios and the acting of the pupils with both classes one week prior to the actual recording. All pupils' legal guardians' written consent was obtained in advance. Before each take, the class including the pupils (i.e. the actresses and actors) received the scripts, a concise briefing and some detailed coaching through the scripted scenarios. The actresses and actors were then instructed to practice their roles in preparation for the following week when the filming took place. Because of this preparation phase, there was enough time during the filming lessons to film retakes and record each scenario in both French *and* English. By recording the same scenario in both languages, I could ensure comparability across the languages and the future research participants' test situation for the pre- and post-test. The following finalised sample script for test task 5 in both English and French encompasses a scenario in which two students (SS1 and SS2) hold a presentation in the L2 classroom on "sport in England" or "sport in France", respectively. In the pre- and post-test, this particular task requires the test takers to watch the scenarios and provide constructive feedback on the presentation to the students depicted in the video-vignette. For creating the

video-vignettes, the pupil actresses and actors received scripts such as the one illustrated below. They were designed in a way that replicates real-world pupil presentation notes:

English script on «Sport in England»	**French script on «Sport in Frankreich»**
SS1: • *Rugby is a popular sport in England and the other countries.* • *There are different kinds of rugby.* • *Rugby union is a sport with two teams with fifteen players.* • *It's the most played kind of rugby.* • *They play with an oval ball in a stadium.* • *The team that has more points than the other team wins the game.* • *They play with the hands for tries, or with the feet for penalties.*	**SS1:** • *Le Rugby est un sport populaire en France et autres pays.* • *Il y a des differentes sortes de rugby.* • *Le rugby union est un sport avec deux équipes et quinze joueurs.* • *C'est la sorte la plus jouée du rugby.* • *Ils jouent avec un ballon ovale dans un stade.* • *L'équipe qui a plus de points que l'autre gagne le jeu.* • *On peut jouer avec les mains pour faire des essais, ou avec les pieds pour des pénalités.*
SS2: • *In rugby league is a very much famous competition.* • *It's called the Six Nations Championship.* • *This competition is between England, France, Ireland, Italy, Scotland and Wales and every year.* • *The champion in 2018 is Ireland.* • *The champion become the "European Champion".* • *England have the record with 28 wins.*	**SS2:** • *Au rugby league, Il y a une compétition beaucoup connue. Elle s'appelle Le Tournoi des Six Nations.* • *Cette compétition est entre Angleterre, France, Irlande, Écosse et pays de Galles et chaque année.* • *Le champion 2018 est Irlande.* • *Le champion devient le «champion d'Europe».* • *Angleterre a le record avec 28 victoires.*

Table 6 : Sample video-vignette scenario script, test task 5

Even though it is one of the main aims of the video- and photo-vignettes to increase the level of authenticity in the test tasks, the participating pupils still "acted". The degree of perceived artificiality could somewhat be compromised through piloting and allowing the actresses and actors to practice their parts, however some of the videos still convey a strong sense of artificiality and awkwardness. To create the photo-vignettes, I used screenshots of selected moments of the video-scenarios, which were pasted into the tasks.

Genuine Text-Material

In addition to video- and photo-vignettes, genuine text-material was to serve as a stimulus in test task 6 to elicit spoken feedback as a task response. In order to achieve this, a colleague researcher and in-service teacher provided an authentic classroom assignment she had implemented earlier in the year, including a genuine text response from a lower-secondary L2 learner:

Exercise
You went on a holiday to London for a week with your parents. In your diary, you are now writing about what you did each day. You are talking about your impressions of life in London. Write a text of 60 to 80 words.

So, Monday, we arrive in London with the plane. Our hotel is called "the king's head» and it is beautiful! My room is big. I have a view on the Tower Bridge of London. Tuesday, we go with the underground to the Buckingham palace and we visit the palace. This was interesting. After we ate dinner in a noble restaurant. Wednesday evening, we walking by the Thames, this is a river. The holidays in London are wonderful!

A comparable task and learner-response was acquired in French. The genuine text samples were then incorporated in the task, in which the test takers were instructed to do the following:

1. Geben Sie eine Rückmeldung zu einem konkreten inhaltlichen Aspekt der schriftlichen Arbeit.
2. Gehen Sie auf zwei gelungene sprachliche Aspekte ein und kommentieren Sie diese.
3. Gehen Sie auf einen sprachlichen Fehler ein und stellen Sie ihn als Lerngelegenheit für Martina dar.

During the test itself, the test takers were then to be provided with the genuine text sample in hard-copy format to closely resemble an authentic real-world situation of a teacher consulting their (genuine) students' writing performances to devise feedback. Like with all other test tasks, task specifications were created throughout the development phase (see above). Figure 13 shows an example of a working task specification created during the development process:

Prüfungsaufgabe 'MÜNDLICHE PRODUKTION: 3.9': Spezifikation Aufgabe 6

Angepasst nach Bachman/Damböck 2018: 101ff.

Unterschiede zwischen realweltlicher Aufgabe und Testaufgabe jeweils hervorheben.

	Realweltliche Aufgabe	**Testaufgaben**
Kurzbeschreibung der Handlung, die in der «Wirklichkeit» bzw. in der Testaufgabe ausgeführt wird (Ausgewählte Deskriptor(en) und/oder Kurzbeschreibung der Handlung als Freitext)	Deskriptor 3.9: «Die Lehrperson kann in der Zielsprache mündliche Rückmeldungen zu Schülerleistungen geben». Die Lehrperson befindet sich in einem Einzelgespräch, in welchem sie eine schriftliche Schülerproduktion mit einer/m Lernenden bespricht. Die Lehrperson studiert die schriftliche Textproduktion der/s Lernenden, macht sich gegebenenfalls Notizen und erteilt eine mündliche Rückmeldung zur schriftlichen Arbeit. Die Lehrperson gibt der/m Lernenden eine mündliche Rückmeldung zu inhaltlichen und sprachlichen Aspekten der schriftlichen Textproduktion.	Deskriptor 3.9: «Die Lehrperson kann in der Zielsprache mündliche Rückmeldungen zu Schülerleistungen geben». Die Testperson erhält eine Beschreibung einer Klassensituation zur Kontextualisierung der Aufgabe. Sie schaut sich eine Bild- und Textvignette (hier: Schülertext) an. Sie muss in der Zielsprache eine kurze Rückmeldung auf eine Schülerleistung (hier: schriftliche Textproduktion) geben (z.B. loben, kritisieren, gratulieren, belohnen, auf Verbesserungsmöglichkeiten hinweisen usw.). Diese Rückmeldung bezieht sich auf inhaltliche und sprachliche Aspekte der Textproduktion.
Konstrukt: Welche (Teil-)**Kompetenzen** werden für die Erfüllung der Aufgabe benötigt (sprachlich, fachdidaktisch, pädagogisch, vorausgesetztes thematisches/Weltwissen, IKT, …)? *[Nicht vergessen: Differenzen markieren]*	Inhaltlich: - Wissen über Spracherwerb und kognitive Entwicklung, um einschätzen zu können, was die Lernenden sprachlich und inhaltlich vom Feedback verstehen - Diagnostische Kompetenzen, um die schriftlichen Sprachkompetenzen von Schüler*innen der Zielstufe Sek 1 einschätzen zu können. - Diagnostische Kompetenzen, um das sprachliche Kompetenzniveau der Schüler*innen einschätzen und die eigenen sprachlichen Produktionen entsprechend gestalten zu können. - Didaktisches und pädagogisches Wissen, um die Wirkung des Feedbacks einschätzen und das Feedback entsprechend gestalten zu können. Sprachlich: - Mündliche Kompetenzen, um adressatengerecht eine Rückmeldung zu inhaltlichen und sprachlichen Aspekten geben zu können (z.B. eigene Wortwahl, grammatische Komplexität, Tempo, Aussprache). - Hohe sprachliche Korrektheit - Klare Aussprache, hohe Flüssigkeit.	Inhaltlich: Genügend Wissen über schriftliche Sprachkompetenzen der Zielstufe, um die beispielhafte Bild- und Textvignette nachvollziehen und darauf reagieren zu können. Sprachlich: - Mündliche Sprachproduktion in der Zielsprache, um den Lernenden basierend auf dem studierten Schülertext eine Rückmeldung zu inhaltlichen und sprachlichen Aspekten (z.B. «Wortschatz», «sprachliche Korrektheit») geben zu können. - Klare Aussprache, hohe Flüssigkeit. - Hohe sprachliche Korrektheit. - Sprachliches Repertoire, das es erlaubt, Wortwahl und grammatische Strukturen an die Kompetenzen der Lernenden anzupassen (z.B. eigene Wortwahl, grammatische Komplexität, Tempo, Aussprache).

Aufgabenmerkmale			
		Realweltliche Aufgabe	**Testaufgaben**
Setting	Raum, Material	Klassenzimmer, Lehrerzimmer, Nebenraum	Computerbasiert, schriftliche Kontextualisierung der Klassensituation, beispielhafte Videovignette zum 1x anschauen, Aufnahmemöglichkeit, ggf. mit Löschfunktion Material: Computer, Kopfhörer, Zugang zu Onlinetest
	Beteiligte (Personenkonstellation: wer kommuniziert mit wem?)	Lehrperson Schüler*innen der Sekundarstufe 1 (7.-9. Klasse)	Rollenspiel / Simulation: Testperson teilt einer/m beispielhaften, in der Bildvignette dargestellten Lernenden etwas mit. Die Antwort wird in ein Mikrofon eingesprochen und online gespeichert und ist monologisch.
	Benötigte Zeit	Vorbereitungszeit: Schülertext studieren – Notizen machen – Rückmeldung basierend auf Kriterien geben, ca.- 10-15 Minuten Abhängig davon, wie lange der Schülertext ist, ca. 20-30 Minuten.	3-5 Minuten Vorbereitungszeit, um die Aufgabe zu studieren und die Schülertext zu analysieren 0.5-2 Minuten Ausführzeit, um die eigene Sprachproduktion aufzunehmen
Input	Form des Inputs (z. B. Audio, Bilder, Fachtext, Lernertext, Items, z. B. Multiple-Choice mit 3 Optionen)	Eingereichte schriftliche Textproduktion von Lernenden Absicht, einer/m Lernenden eine Rückmeldung zu ihren Sprachkompetenzen zu geben	Online Testaufgabe Kontextsetzung und Aufgabeninstruktionen Beispielhafte Videovignette Anweisung, einer/m Lernenden eine Rückmeldung zu ihren schriftlichen Sprachkompetenzen zu geben
	Merkmale der Sprache im Input (z. B. Komplexität, Wortschatz)	Eingereichte schriftliche Textproduktion von Lernenden. Abhängig von der schriftlichen Textproduktion der Lernenden.	Eingereichte schriftliche Textproduktion von Lernenden. Textvignette: sprachlicher Input in Form einer schriftlichen Textproduktion von Lernenden in der Zielsprache.
	Länge (z. B. Wortanzahl, Dauer)	Abhängig von der Länge der schriftlichen Textproduktion, ca. 10-20 Minuten.	Bild- und Textvignette: ca. 7 Minuten
	Erlaubte Themen	Einzelgespräch mit kurzer Rückmeldung auf schriftliche Textproduktionen von Lernenden: Prinzipiell jegliche Art von schriftlichen Schülertexten, welche in einem Klassenzimmer denkbar sind.	Einzelgespräch mit kurzer Rückmeldung auf schriftliche Textproduktionen von Lernenden: Prinzipiell jegliche Art von schriftlichen Schülertexten, welche in einem Klassenzimmer denkbar sind. Textvignette: schriftliche Textproduktion zu einem von der LP im Voraus bestimmten Thema.
Erwarteter Output (Leistung, Antwort)	Form (z. B. mündliche Erklärung für Lernende, mündliche Erzählung, schriftliche Anweisung, korrigierte Fehler)	Mündliches Feedback durch die Lehrperson, angepasst an die schriftliche Textproduktion und die Sprachkompetenz der Lernenden.	Aufnahme eines mündlichen Feedbacks durch die Testperson in der Zielsprache. Das mündliche Feedback geht auf die in der Textvignette dargestellte schriftliche Textproduktion von Lernenden ein und ist an die Sprachkompetenz der Lernenden angepasst.

Figure 12 : Excerpt working-task-specification task 6

After filming and editing the clips, I made a few more alterations to the task instructions and task specifications to ensure the video-vignettes matched the scenarios accordingly. I then

incorporated the video-vignettes in the respective test tasks. In this beta test-version, each of the seven test tasks contained the following:

- an estimated task completion time (depending on the complexity of the task, the completion time ranged between 4 and 10 minutes),
- the scope of the task (between 30 seconds and 2 minutes of active speaking time),
- a description of the target audience which the participants needed to tailor their language production to (lower-secondary school students at different L2 proficiency levels),
- a detailed description of the scenario and context of each task,
- a photo-, video- or genuine text-vignette that provides a necessary real-world stimulus, and
- precise test task instructions.

Subsequently, this beta version of the test could be implemented in the chosen learning management system (LMS) be pre-piloted and piloted respectively. In the following section, I outline the specifics of the chosen LMS as an online test environment and the implementation of the test tasks.

4.3.2.2. Moodle Implementation

With the decision that this test was to be computer-based and online administered, a suitable online environment needed to be determined. Moodle (Moodle, 2021) is a powerful and versatile tool (Douglas, 2010) that can be used for designing and delivering online classes and assessment modules. It is suitable for test development projects as it allows for various task types such as, among others, quizzes, matching, multiple choice, true/false, a type of cloze, and short answer or long answer and essay tasks (Douglas, 2010). It is freely accessible to all PHSG staff and students through their individual PHSG log-ins. It allows for uploading and integrating audio and video files in test tasks via SWITCHtube[21], and most importantly, it contains a microphone option that enables test takers to record and save their speech productions directly on the platform. Finally, downloading and storing test task responses is straightforward. The PHSG-predecessor of the present test (see above) was implemented in and similar data

[21] SWITCHtube is a service offered by the SWITCH Foundation (SWITCH Information Technology Services; the Swiss national research and education network organisation). It constitutes an online platform that enables "academic video sharing". See https://tube.switch.ch/ for more information (last accessed: 16.6.2021)

collections were conducted via Moodle prior to this study. During these preceding efforts, Moodle proved to be of high usability and hence to suit the present test purpose.

Once all test tasks were completed, the tasks including all stimuli were migrated to Moodle in preparation for the pre-piloting and piloting phase of the test. The following figure illustrates the task layout with a video-vignette sample task including the voice-recording function (circled):

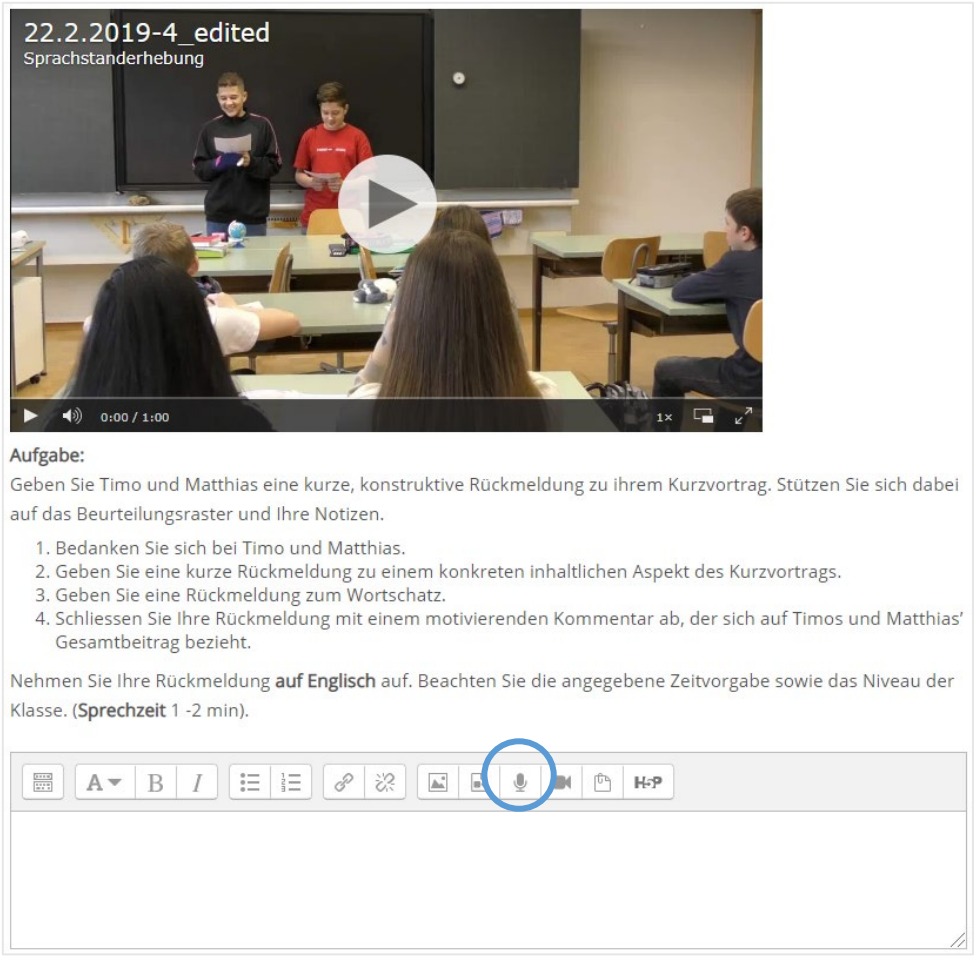


Figure 13 : Sample test task embedded in Moodle

4.3.2.3. Piloting and Revision of Test Items

Once all preliminary test items and video-, photo- and text-vignettes were migrated to Moodle, the test was ready for piloting. In a pre-pilot study, language testing and language teaching experts scrutinised the test items throughout the test item construction process in several auditing cycles to provide qualitative expert judgements on the content- and construct validity of the test items. After their feedback was incorporated and the feedback cycle reached a perceived level of saturation, the test was piloted with three undergraduate student assistants

who at the same time represented the test taker target population. I conducted the piloting phase employing the think-aloud method (Knoblich & Öllinger, 2006; Weidle & Wagner, 1982). This research method involves participants articulating their thoughts while completing a specific task, thereby enabling the researcher to gain insight into the participants' informative cognitive processes that occur during specific actions (Sandmann, 2014; Weidle & Wagner, 1982). In addition, especially in test validation, it allows for examining the quality of test items through identifying the comprehensibility of the tasks or problems thereof (ibid.). Think-aloud protocols are considered a useful method for assessing test quality criteria and for validating test items (Weidle & Wagner, 1982). I audio-recorded the think-aloud pilot sessions and took notes during the process. Immediately after the test completion, I asked the pilot study participants more clarifying questions about their actions, reactions and articulations during the test completion. The sessions took place sequentially, which allowed adaptations based on the insights gained from participant 1 in preparation for the pilot study with participant 2 and participant 3 respectively. Some of the changes implemented included more specific contextualisation and clearer instructions for each test item and the addition of a warm-up test task in the target language to help the test takers adjust to the test situation. This iterative process therefore enabled the improvement of the test instrument throughout the actual process, which resulted in only minimal additional insights and necessary changes after participant 3 completed the task. The final test structure looked as follows and included the following test tasks:

- A welcome-landing page including general information about the data collection,
- an online form of consent for participants,
- general instructions on how to work through the test and record the task responses,
- the following test tasks:
    - An introductory warm-up task (see above)
    - Task 1: Provide oral feedback to a learner's input and encourage class participation (descriptor 3.7)
    - Task 2: Provide oral feedback to the overall behaviour of a group of learners (descriptor 3.7)
    - Task 3: Conduct a dialogue with a learner in order to give feedback on her ability express themselves in a particular TLU context in an appropriate and grammatically accurate manner based on PRLC-R criteria (descriptor 3.8a)

  - Task 4: Provide a learner with feedback on her ability to participate in a conversation in the target language based on PRLC-R criteria (descriptor 3.7a)
  - Task 5: Provide oral feedback on a learning group's oral presentation (descriptor 3.9)
  - Task 6: Provide feedback on a learner's written performance (descriptor 3.9)
  - Task 7: Hold an advisory talk with an English native-speaker on their strengths and weaknesses with reference to classroom participation with the aim of fostering their skills in a personalised manner (descriptor 3.12)
- A final page with instructions on how to submit the test responses and a thank-you message.

After the insights from the pilot study were compiled and the final revision of the test items was completed, the test was ready to be used in the actual data collection (see chapters 4.4.3 and 4.4.5). The next section introduces the detailed main-study design including the test administration procedures.

4.4. Design Main-Study

As outlined above, the BA E-Portfolio task B provided the environment for the present intervention study. In the original task B, students provided peer feedback without receiving guidance or feedback training (see chapter 4.1). The way students structured and provided their feedback was not uniformly regulated. Indeed, the feedback instructions largely depended on the preferences of the individual learning group's supervisor and ranged from not being explicitly requested to requiring students to define their own evaluation criteria on which they based their peer feedback. This particular component of the BA E-Portfolio task B consequently led to ample criticism from the project group supervisors, as they equally perceived constraints and a lack of quality in students providing concrete, useful and effective feedback. These perceptions align with findings from the literature regarding students' tendency to provide superficial feedback (Leki, 1990; Lockhart & Ng, 1993; Mendonca & Johnson, 1994; Min, 2016; Tsui & Ng, 2000). To investigate RQ #1, the peer-feedback component of the original BA E-Portfolio task B format was manipulated. Instead of providing unguided and untrained peer feedback, and in close alignment to the research design of Yeh et al. (2019), the treatment

involved the application of specific evaluation criteria retrieved from the PRCL-R. The benefits of rubrics in a learning environment are manifold (see chapter 1.3). Thus, the PRLC-R offered a suitable tool for peer feedback to raise awareness and promote the development of the oral profession-related language competences of AoA 3 and teacher and student feedback literacy. The application of the PRLC-R as a treatment served two purposes:

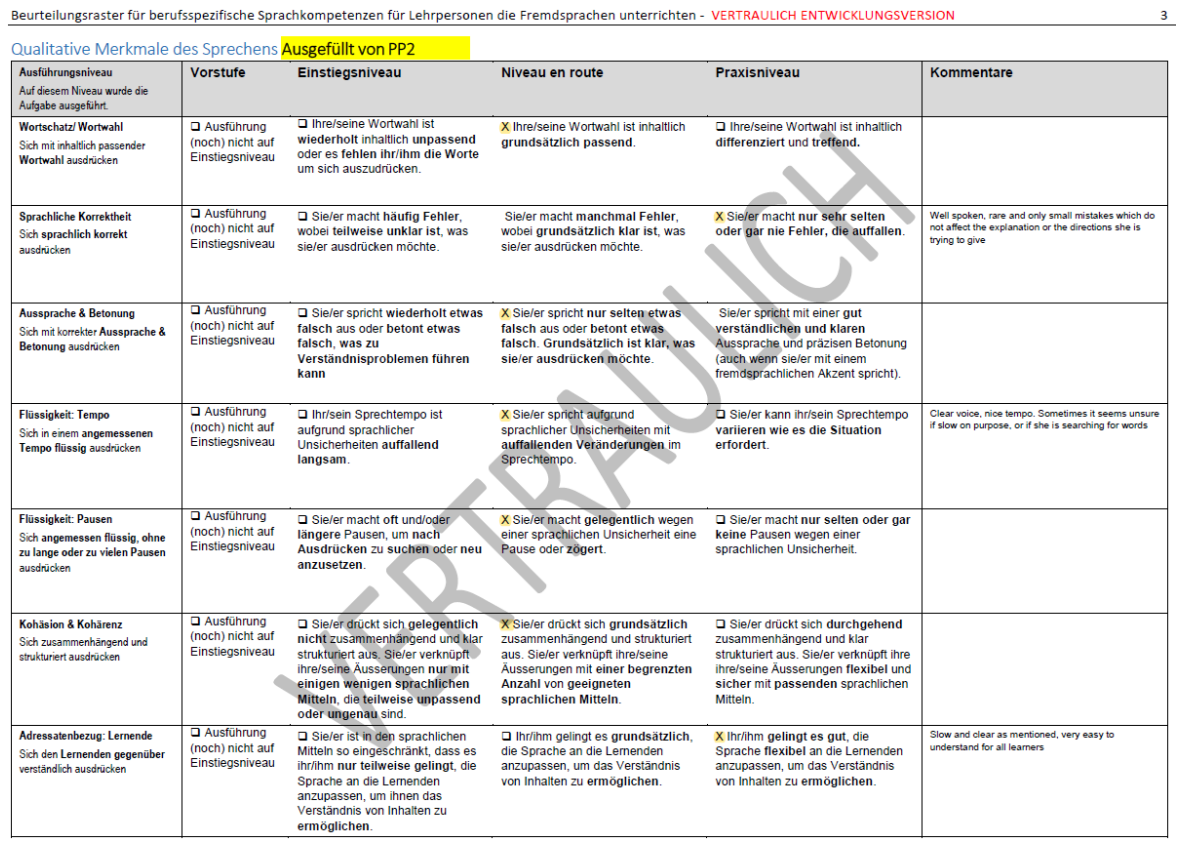
- to provide the basis for the peer feedback and guide students' comments along explicit criteria specifically designed to evaluate profession-related language competences. Therefore, the PRLC-R serve to enable students to provide more guided, focused, and perhaps higher quality peer feedback,
- to investigate the development of language-specific aspects of peer feedback throughout the E-Portfolio process (RQ #1).

Because rubrics are not self-explanatory and their use needs to be trained (Birri & Smit, 2013), the treatment included a PRLC-R training component at the commencement of the intervention (see chapter 4.4.4). This updated multi-stage assignment thus allowed students to work iteratively and explicitly on their oral, profession-related language competences through the combination of language development and feedback practice. In order to be able to identify any treatment effects, I devised an experimental-control-group design where the 50 research participants were divided into three comparison groups. The phil I students were subdivided into an experimental group E (n=21) and a control group C1 (n=19), and the phil II students constituted a second control group C0 (n=10) who, in contrast to E and C1, did not complete the BA E-Portfolio (see chapter 4.2). Variables such as gender and language focus subject (English, French, or a combination of the two) were controlled for. Out of the 40 participants who completed the BA E-Portfolio (E and C1), six students devised their portfolio in French. The remaining 34 completed their portfolios in English. By including all students of the complete cohort in the study, I aimed to ensure that none of the participants would feel "punished" or "rewarded" for studying one or the other language, and that instead they felt included in the cohort as a whole. Because this research focuses on the potential effects of the PRCLP and PRLC-R on students' English oral profession related language competences, data from the French students' pre- and post-test responses were gathered but excluded from the data analysis. All students participated in the pre-test, and one student from the English cohort had dropped out by the time of the post-test, leaving a sample of 33 participants whose data could be used for analysis. Before the implementation of the treatment, the application of the

PRLC-R in the relevant context needed to be piloted. I will proceed to describe the piloting phase in the next section.

4.4.1. The Pilot Study

To verify the feasibility of including and applying the PRLC-R as a central component in the BA E-Portfolio assignment, I conducted an informal consultation with L2 teacher educators and a pilot study. In the informal consultation, language teaching and learning experts from the PHSG applied the PRLC-R to assess sample oral L2 productions of PHSG students. Aside from constituting an initial exploration of the PRLC-R usability in context, the consultation served to gain insights to then further refine the rubric for its subsequent implementation in the pilot study. The pilot then tracked two BA E-Portfolio project groups' dealings with the PRLC-R over the course of their entire e-portfolio assignment in the academic year 2018-2019. At this point, the PRLC-R contained the following assessment criteria: *Wortschatz/Wortwahl, Sprachliche Korrektheit, Aussprache & Betonung, Flüssigkeit: Tempo, Flüssigkeit: Pausen, Kohäsion & Kohärenz* and *Adressatenbezug: Lernende.* In addition, the rubric encompassed four performance level labels (PLLs) termed *Vorstufe*, *Einstiegsniveau*, *Niveau en route* and *Praxisniveau* with a PLD for each criterion at each performance level (see Figure 14 below). The comments box allowed participants to specify their evaluation. Prior to the commencement of the student feedback cycles of the BA E-Portfolio 2018-2019, one group of two and one group of four pilot-study participants (PP) were recruited, familiarised with the rubric in a face-to-face introductory session, and instructed with regard to its application. Special emphasis was put on the comments box to encourage the participants to extensively comment on their judgements. The following excerpt from the piloting phase shows an example of how pilot-study participant 2 (PP2) applied the PRLC-R:



Beurteilungsraster für berufsspezifische Sprachkompetenzen für Lehrpersonen die Fremdsprachen unterrichten - VERTRAULICH ENTWICKLUNGSVERSION 3

Qualitative Merkmale des Sprechens Ausgefüllt von PP2

Ausführungsniveau Auf diesem Niveau wurde die Aufgabe ausgeführt.	Vorstufe	Einstiegsniveau	Niveau en route	Praxisniveau	Kommentare
Wortschatz/ Wortwahl Sich mit inhaltlich passender **Wortwahl** ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Ihre/seine Wortwahl ist **wiederholt** inhaltlich **unpassend** oder es **fehlen ihr/ihm die Worte** um sich auszudrücken.	X Ihre/seine Wortwahl ist inhaltlich **grundsätzlich passend**.	❑ Ihre/seine Wortwahl ist inhaltlich **differenziert** und **treffend.**	
Sprachliche Korrektheit Sich **sprachlich korrekt** ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Sie/er macht **häufig Fehler**, wobei **teilweise unklar ist**, was sie/er ausdrücken möchte.	Sie/er macht **manchmal Fehler**, wobei **grundsätzlich klar ist**, was sie/er ausdrücken möchte.	X Sie/er macht **nur sehr selten oder gar nie Fehler, die auffallen**.	Well spoken, rare and only small mistakes which do not affect the explanation or the directions she is trying to give
Aussprache & Betonung Sich mit korrekter **Aussprache & Betonung** ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Sie/er spricht **wiederholt etwas falsch** aus oder **betont etwas falsch, was zu Verständnisproblemen führen kann**	X Sie/er spricht **nur selten etwas falsch** aus oder **betont etwas falsch. Grundsätzlich ist klar, was sie/er ausdrücken möchte**.	Sie/er spricht mit einer **gut verständlichen und klaren** Aussprache und präzisen Betonung (auch wenn sie/er mit einem fremdsprachlichen Akzent spricht).	
Flüssigkeit: Tempo Sich in einem **angemessenen Tempo flüssig** ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Ihr/sein Sprechtempo ist aufgrund sprachlicher Unsicherheiten **auffallend langsam**.	X Sie/er spricht aufgrund sprachlicher Unsicherheiten mit **auffallenden Veränderungen** im Sprechtempo.	❑ Sie/er kann ihr/sein Sprechtempo **variieren wie es die Situation erfordert**.	Clear voice, nice tempo. Sometimes it seems unsure if slow on purpose, or if she is searching for words
Flüssigkeit: Pausen Sich **angemessen flüssig, ohne zu lange oder zu vielen Pausen** ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Sie/er macht **oft** und/oder **längere** Pausen, um **nach Ausdrücken** zu **suchen** oder **neu anzusetzen**.	X Sie/er macht **gelegentlich** wegen einer sprachlichen Unsicherheit eine Pause oder **zögert**.	❑ Sie/er macht **nur selten oder gar keine** Pausen wegen einer sprachlichen Unsicherheit.	
Kohäsion & Kohärenz Sich zusammenhängend und strukturiert ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Sie/er drückt sich **gelegentlich nicht** zusammenhängend und klar strukturiert aus. Sie/er verknüpft ihre/seine Äusserungen **nur mit einigen wenigen sprachlichen Mitteln**, die **teilweise unpassend oder ungenau** sind.	X Sie/er drückt sich **grundsätzlich** zusammenhängend und strukturiert aus. Sie/er verknüpft ihre/seine Äusserungen mit **einer begrenzten Anzahl** von **geeigneten sprachlichen Mitteln**.	❑ Sie/er drückt sich **durchgehend** zusammenhängend und klar strukturiert aus. Sie/er verknüpft ihre ihre/seine Äusserungen **flexibel** und **sicher** mit **passenden** sprachlichen Mitteln.	
Adressatenbezug: Lernende Sich den **Lernenden gegenüber** verständlich ausdrücken	❑ Ausführung (noch) nicht auf Einstiegsniveau	❑ Sie/er ist in den sprachlichen Mitteln so eingeschränkt, dass es ihr/ihm **nur teilweise gelingt**, die Sprache an die Lernenden anzupassen, um ihnen das Verständnis von Inhalten zu **ermöglichen**.	❑ Ihr/ihm gelingt es **grundsätzlich**, die Sprache an die Lernenden anzupassen, um das Verständnis von Inhalten zu **ermöglichen**.	X Ihr/ihm **gelingt es gut**, die Sprache **flexibel** an die Lernenden anzupassen, um das Verständnis von Inhalten zu **ermöglichen**.	Slow and clear as mentioned, very easy to understand for all learners

Figure 14 : PRLC-R version for the pilot study

After the PPs had submitted their ratings and completed their e-portfolios, they provided written feedback on the usability, application and comprehensibility of the rubric. The main insights gained from the pilot study related to the (in)comprehensibility and fuzziness of individual criteria and some specific terminology. For example, the participants found it challenging to distinguish between fluency (*Flüssigkeit: Pausen*) and articulation/speech rate (*Flüssigkeit: Tempo*) and found themselves providing redundant or repetitive feedback in both categories. After carefully reconsidering the affordances and challenges of both criteria to an accurate assessment of profession-related language competences, I removed the criterion articulation rate (*Flüssigkeit: Tempo*) from the PRLC-R. In addition, students found some of the terms in the PRLC-R unclear. For example, they found it difficult to understand what *Sprachliche Mittel* (linguistic competences) entailed. In order to clarify this concept, a footnote with a definition from the CEFR (Council of Europe, 2001) was added to the PRLC-R. Finally, the participants exhibited a strong tendency towards the center when applying the PRLC-R, primarily choosing the performance level *Niveau en route* when evaluating their peers. Additionally the participants reported insecurities and difficulties when interpreting the meaning of the individual levels and when deciding which performance level to assign their peers to, as illustrated in the following quote:

> Like in every assessment grid, there sometimes is a bit of uncertainty about which level to put the peer in. Meaning because it is a grid, somebody is for example, either bad, not bad, good or very good. This is not specifically about this grid, but about using grids in general There is sometimes a lack of options. Therefore it is important to have a column open for the comments, that way one can specify her or his chosen category. (sic. PP2)

There was some indication that the PLLs caused too much room for interpretation and thus lead to more confusion rather than clarification. As a result, I replaced the PLLs with the numbers *0, 1, 2* and *3*. To counteract student evaluators' tendency to the center and to respond to the participants' requests for more options, I subdivided each performance level into two sub-levels. Due to the otherwise positive feedback regarding the usability, application, clarity and benefits of the PRLC-R with reference to their individual L2 development process, I did not perform any further modifications. After these modifications, the PRLC-R looked as follows:

Fremdbeurteilungsraster: Qualitätsmerkmale des Sprechens

Beurteiler*in: _____ Feedback-Empfänger*in: _____ Unterrichtssequenz Nr.: _____

Ausführungsniveau Qualität der Aufgabenausführung	**0**	**1**	**2**	**3**	**Kommentare**
Wortschatz / Wortwahl Sich mit inhaltlich passender **Wortwahl** ausdrücken	❑ Ausführung (noch) nicht wie 1	Ihre/seine Wortwahl ist **wiederholt** inhaltlich **unpassend** oder es fehlen ihr/ihm die Worte um sich auszudrücken.	Ihre/seine Wortwahl ist inhaltlich **grundsätzlich passend**.	Ihre/seine Wortwahl ist inhaltlich **differenziert** und **treffend.**	
		❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	
Sprachliche Korrektheit Sich **sprachlich korrekt** ausdrücken (z.B. Grammatik)	❑ Ausführung (noch) nicht wie 1	Sie/er macht **häufig** Fehler, wobei **teilweise unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **manchmal** Fehler, wobei **grundsätzlich klar** ist, was sie/er ausdrücken möchte.	Sie/er macht **nur sehr selten oder gar nie** Fehler, die auffallen.	
		❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	
Aussprache & Betonung Sich mit korrekter **Aussprache & Betonung** ausdrücken	❑ Ausführung (noch) nicht wie 1	Sie/er spricht **wiederholt** etwas falsch aus oder betont etwas falsch, was zu **Verständnisproblemen** führen kann	Sie/er spricht **nur selten etwas falsch** aus oder betont etwas falsch. **Grundsätzlich ist klar**, was sie/er ausdrücken möchte.	Sie/er spricht mit einer **gut verständlichen** und klaren Aussprache und **präzisen** Betonung (auch wenn sie/er mit einem fremdsprachlichen Akzent spricht).	
		❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	
Flüssigkeit Sich in einem **angemessenen Tempo flüssig** ausdrücken	❑ Ausführung (noch) nicht wie 1	Ihr/sein Sprechtempo ist aufgrund sprachlicher Unsicherheiten **auffallend langsam**.	Sie/er spricht aufgrund sprachlicher Unsicherheiten mit **auffallenden Veränderungen** im Sprechtempo.	Sie/er kann ihr/sein Sprechtempo **variieren** wie es die Situation **erfordert**.	
		❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	
Kohäsion & Kohärenz Sich zusammenhängend und strukturiert ausdrücken	❑ Ausführung (noch) nicht wie 1	Sie/er drückt sich **gelegentlich nicht zusammenhängend** und **nicht klar strukturiert** aus. Sie/er verknüpft ihre/seine Äusserungen nur mit einigen wenigen sprachlichen Mitteln*, die **teilweise unpassend** oder ungenau sind.	Sie/er drückt sich **grundsätzlich zusammenhängend** und **strukturiert** aus. Sie/er verknüpft ihre/seine Äusserungen mit **einer begrenzten** Anzahl von **geeigneten** sprachlichen Mitteln*.	Sie/er drückt sich **durchgehend zusammenhängend** und **klar strukturiert** aus. Sie/er verknüpft ihre ihre/seine Äusserungen **flexibel** und **sicher** mit **passenden** sprachlichen Mitteln*.	
		❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	
Adressatenbezug Sich den **Lernenden gegenüber** verständlich ausdrücken	❑ Ausführung (noch) nicht wie 1	Sie/er ist in den sprachlichen Mitteln* so **eingeschränkt**, dass es ihr/ihm **nur teilweise** gelingt, die Sprache an die Lernenden anzupassen, um ihnen das Verständnis von Inhalten zu ermöglichen.	Ihr/ihm **gelingt es grundsätzlich**, die Sprache an die Lernenden anzupassen, um das **Verständnis** von Inhalten zu ermöglichen.	Ihr/ihm **gelingt es gut**, die Sprache flexibel an die Lernenden anzupassen, um das Verständnis von Inhalten zu ermöglichen.	
		❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	❑ trifft eher zu ❑ trifft zu	

Figure 15: Finalised version of the PRLC-R for the intervention study

The insights from the pilot study also contributed to the refinement of the design of the main-study. For example, to channel each participant's focus (as well as for practical and theoretical reasons, see chapter 2.4.3), I decided that in the main-study the participants would work with a designated feedback partner. This limitation of the amount of people they would provide

feedback to and receive feedback from to only one participant would thus encourage dialogic feedback and the development of (student) feedback literacy. Apart from including the PRLC-R as a foundation for feedback, the peer feedback process during the pilot study was still relatively open in terms of how students decided to implement that part of the task. To create an environment that simulates more closely an authentic feedback situation in the L2 classroom, I decided to implement a more clearly structured peer feedback process in the intervention. For the treatment, students would be required to provide their peer feedback in a face-to-face discussion immediately following the microteaching sequence. The temporal proximity of conducting the microteaching sequence and providing and receiving feedback respectively represents an attempt to increase the likelihood for students to produce and practise near-authentic spoken language in form of feedback. Implementing these modifications served as the final step to finalise the preparations for the intervention. To understand the entire intervention design in detail, I outline the individual steps in the subsequent section.

4.4.2. Treatment

With the course *Introduction to Linguistics* introducing and facilitating the BA E-Portfolio (see chapter 4.1), its first lecture of the 2019 spring term constituted the beginning of the actual intervention study. Prior to the opening lecture of the course, the E and C1 groups received an informative e-mail from the course convener with the most important milestones of the BA E-Portfolio. Simultaneously, they were instructed to choose the language in which they were going to devise the BA E-Portfolio Part B (English, French or German). Based on the language they chose, the participants formed learning groups of three to four students in preparation for the opening lecture. At the same time, the entire cohort (E, C1 and C0) was informed about PHSG-internal quality control procedures connected to the present dissertation study. The informative e-mail included a description and rationale of said quality measures, including the aim of the measures to evaluate the appropriateness and effectiveness of 2019/2020 *Introduction to Linguistics* course and the BA E-Portfolio assessment. As part of this project, the cohort was also informed that all students (E, C1 and C0) were going to undertake a language competence test (pre-test) prior to commencing and after completing (post-test) the BA E-Portfolio to evaluate the effectiveness of both the course and BA E-Portfolio. After the participants had formed the project learning groups, and prior to the opening lecture, I randomly allocated each learning group to either the E or C1 group. Thus, randomisation was executed on the group level rather than the individual, personal level. Additionally, the two French

project groups (six participants) were distributed evenly across the E and C1 group (one learning group per treatment each). The following figure shows the structure of the main-study:

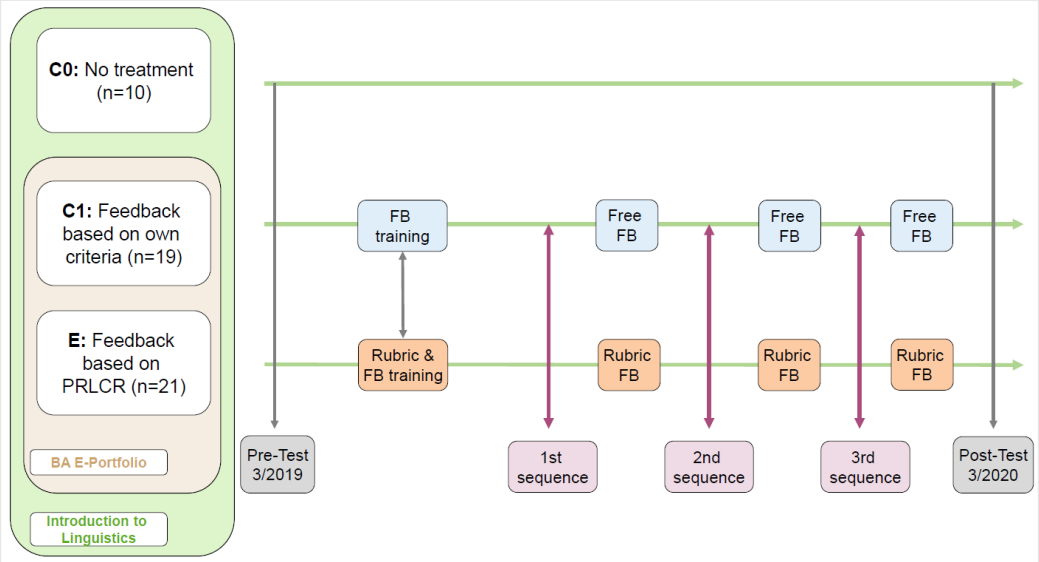


Figure 16 : Outline of the intervention design of the main-study

As introduced above, the comparison groups differed from each other with reference to how they were to complete the BA E-Portfolio assignment. While C0 did not complete the BA E-Portfolio task but solely partook in the pre- and post-test, the E and C1 groups underwent the entire assignment process. The E group devised the BA E-Portfolio based on the PRLC-R, and the C1 group did not receive any pre-defined assessment criteria but developed their own within the respective learning groups (see chapter 4.3.). Similar to the original format (see chapter 4.1), the participants completed three cycles of the following three steps: 1) devising and video-recording a microteaching sequence, 2) providing (dialogic) peer feedback on their designated partners' microteaching sequence (cf. Ajjawi & Boud, 2018; Carless et al., 2011), and 3) reflecting on and optimising their microteaching sequence based on the received feedback. The peer feedback component of step 2) involved participants providing their peer feedback in a face-to-face discussion (cf. Ajjawi & Boud, 2018; Carless et al., 2011) immediately following the respective microteaching sequence in order to simulate an environment that closely corresponds to an authentic feedback situation. This step also served to create an environment conducive to the development of feedback literacy. The temporal proximity of the microteaching sequence and the feedback process attempted to increase the likelihood for students to produce near-authentic spoken language in form of feedback. It should also reduce the likelihood for them to write their feedback in prose and read it aloud – a step that would jeopardise the spontaneity of speaking and compromise possible training effects. By giving and

receiving feedback from a learning group member in dialogic feedback conversations during each of the three cycles, students had the opportunity to implicitly train their feedback and advising skills in the target language and develop their feedback literacy. These competences correspond with the descriptors of AoA 3, and students arrived at full circle: upon completion of the BA E-Portfolio part B, they had trained their oral, profession-related language skills with a focus on giving feedback. Based on the conceptual framework of teacher and student feedback literacy (Carless, 2020a, 2020b; Carless & Boud, 2018; Carless et al., 2011; Chong, 2021), the participants fulfilled a double role in this intervention study. On the one hand, through completing the BA E-Portfolio part B, they participated in an activity that fostered the development of student feedback literacy, communicative competence and evaluative judgement. On the other hand, the participants' iterative participation in dialogic peer feedback served as an attempt to circumvent the limitations of the unidirectional transmission of feedback and to reconcile teachers' and students' (here: feedback-provider and feedback-recipient) differing views and understandings of feedback. This practice also enabled the research participants to train their language analytical ability to promote student comprehension, feedback uptake and teacher feedback literacy (Shintani & Ellis, 2015). The following table outlines the treatment steps for the comparison groups in chronological order:

Action step	**E**	**C1**	**C0**
Orientation about study and pre- and post-test			
Choosing language to devise the BA E-Portfolio part B	✓	✓	✓
Forming learning groups	✓	✓	
Allocating learning groups to E and C1			
Opening lecture *Introduction to Linguistics*			
Pre-test	✓	✓	✓
Introduction to BA E-Portfolio	Introduction to task & mention of PRLC-R	Introduction to task & mention of devising own assessment criteria	
PRLC-R and feedback training			
	PRLC-R & feedback training	Devising own assessment criteria & feedback training	
Devising BA E-Portfolio task B			
1st microteaching	✓	✓	
1st peer feedback	PRLC-R & feedback dialogue	Own criteria & feedback dialogue	
Feedback reflection for 2nd microteaching	✓	✓	

2nd microteaching	✓	✓	
2nd peer feedback	PRLC-R & feedback dialogue	Own criteria & feedback dialogue	
Feedback reflection for 3rd microteaching	✓	✓	
3rd microteaching	✓	✓	
3rd peer-feedback	PRLC-R & feedback dialogue	Own criteria & feedback dialogue	
Conclusion of treatment			
Post-Test	✓	✓	✓
BA E-Portfolio submission	✓	✓	

Table 7 : Outline of the individual action steps of the main-study intervention

Before outlining the data processing and test scoring procedures, the administration of the pre- and post-test as well as the feedback and PRLC-R training are discussed in the following subchapter.

4.4.3. Pre-Test

The administration of any test requires the consideration of elements that can potentially compromise the test's reliability and lead to erroneous interpretations of test takers' performance (Douglas, 2010). While some elements are beyond the control of the test administrators and test takers, aspects such as the test environment, the personnel, the procedures and the scoring need to be carefully considered and controlled (ibid.). Prior to the administration of the pre-test, the research participants received instructions including a virtual guidance document that served to familiarise them with the test purpose and the test format. With the pre- and post-test being a computer-based, online-administered test, the test takers were instructed to bring their own devices to the test. In order to avoid any technical problems on the test-day and to become familiar with the test tasks, the research participants received access to a mock test on Moodle they needed to complete prior to the actual test. This trial run consisted of a general introduction to the test and two mock tasks similar to those of the actual test. The pre-test was then administered to the entire cohort (E, C1 and C0) during the opening lecture of the course *Introduction to Linguistics* in March 2019. As mentioned above, the opening lecture simultaneously served the formal introduction to the BA E-Portfolio task. Since the BA E-Portfolio assignment differed from the experimental group E and the control group C1, and since it was crucial that both groups remained unaware of these differences, group E

and group C1 were separated from each other for the pre-test and the formal BA E-Portfolio introduction. During the first half of the opening lecture, the participants of the control group C were introduced to the BA E-Portfolio task by the course convener while the experimental group E completed the pre-test in separate rooms with three to a maximum of five students per room. It was important for the number of students not to exceed five per room to mitigate the risk of test takers either disturbing or influencing one another during the recording of their test responses. This risk could naturally not be eliminated completely, however it was mitigated to the extent to which it could be ensured that the test was maximally useful in providing the necessary information about the test takers' abilities (Bachman & Palmer, 1996). These measures also allowed for the administration and supervision of the test to remain within what was feasible. The complexity of the entire test administration required twelve test supervisors, all of whom received training and careful written instructions prior to the test-day as well as verbal explanations with the opportunity to ask questions immediately before the test administration (Douglas, 2010). The pre-test took around 30 minutes and a test administration assistant supervised the test participants in each room. During the test itself, all supervisors could reach out to one another via mobile phone in case that further assistance was needed. During the second half of the opening lecture, the roles of the groups were reversed. While the control group C1 and the control group C0 (who was told to only arrive for the second half of the lecture) completed the same pre-test under identical conditions to the experimental group E, the experimental group E was formally introduced to the BA E-Portfolio task. For ease of understanding, the below working document depicts the pre-test process and organisation:

Gruppen-, Zeit- und Raumeinteilung

10:35-11:20: Einführung E (Aula), Prä-test K & Nullgruppe Phil II

11:25-12:10: Einführung K (Aula), Prä-test E Nullgruppe Phil II

Nicht markiert: Englisch Test; Pink markiert: Französisch Test

Helfer (H)	**Räume**	**10:35-11:20** Einführung E Test K	**11:25-12:10** Einführung K Test E & Phil II
H1	Aula	(**20** TN)	(**30** TN)
H2	G203	TN1 TN2 TN3	TN21 TN22 TN23
H3 & H4	G155	**H3:** TN4 TN5 TN6	**H4:** TN24 TN25 TN26
H5	G042	TN7 TN8 TN9	TN27 TN28 TN29
H6	G105	TN10 TN11 TN12	TN30 TN31 TN32
H7	G231	TN13 TN14 TN15	TN33 TN34 TN35
H8	G205	TN16 TN17 TN18	TN36 TN37 TN39
H9	G110	TN19 TN20	TN39 TN40
H10 Phil II	G240	-	TN41 TN42 TN43 TN44 TN45
H12 Phil II	G241	-	TN46 TN47 TN48 TN49 TN50

Figure 17 : Organisation and process of pre-test administration

Key

E	Experimental group E	Phil II	Control group C0
H	Helper /assistants	TN	Research participant
K	Control group C1		

The supervisors collected all additional test materials after the test completion, which were safely stored for data analysis. Additionally, all data were downloaded from Moodle,

anonymised and safely stored on an external storage device immediately after the pre-test administration (Douglas, 2010).

4.4.4. Feedback and Rubric Training

When assigning students to work with a rubric, a comprehensive introduction to its use and application is crucial (Cheung-Blunden & Khan, 2018). In addition, providing effective feedback is not a skill that develops over time but a skill that specifically needs to be learned and practiced. Therefore, an additional feedback and rubric training session was scheduled for the experimental group E and the control group C1 in the autumn term 2019, just before the learning groups started providing peer feedback on their first video-recorded multimedia sequences. The duration of an entire lecture was set aside for these training purposes and, much like for the pre-test, the learning groups of group E and of group C1 were separated from each other in order to ensure that the differences in task remained concealed to the research participants. They were also separated from each other with reference to their language focus. These administrative constraints lead to the formation of four separate groups (E group English, E group French, C1 group English and C1 group French). To minimise any possible (novelty) bias or halo effects, four project group supervisors simultaneously conducted the training on in separate rooms. The following table illustrates the organisation of the procedure in more detail:

Rooms	Feedback training	Assessment criteria	Practice
E group English	Best practice feedback principles	Introduction to PRLC-R	Practice application of PRLC-R with 2 exemplars
E group French	Best practice feedback principles	Introduction to PRLC-R	Practice application of PRLC-R with 2 exemplars
C1 group English	Best practice feedback principles	Creating own criteria in learning groups	Practice application of own criteria with 2 exemplars
C1 group French	Best practice feedback principles	Creating own criteria in learning groups	Practice application of own criteria with 2 exemplars

Table 8 : Overview feedback and rubric training

At the beginning of each individual training session, the conveners familiarised all research participants equally with best practice principles of providing constructive feedback (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006). Group E was then introduced to the PRLC-R and the feedback process they were expected to follow. In order to make the criteria relatable and understandable, they worked with two exemplars of student performance for training

purposes. The exemplars were retrieved from the students who completed the BA E-Portfolio in a previous year. Thus, the exemplars were context-specific and context-relevant for the research participants. Working with such benchmarks is known to foster a mutual understanding of what is considered a "good" performance (Bacchus et al., 2020; McNamara, 1996; North, 2000). In contrast to receiving pre-defined assessment criteria extracted from the PRLCR, group C1 identified relevant assessment criteria within their respective learning groups and tested them on the same two exemplars. Both group E and group C1 received opportunities to ask questions to become confident with applying the PRLC-R and their individually chosen assessment criteria respectively.

4.4.5. Post-Test

One year later, just after the learning groups provided their last round of oral feedback and just before they submitted their individual BA E-Portfolios, the entire 2017-2022 cohort (E, C1 and C0) completed the post-test. In order to ensure comparability of the test responses, to enable a reliable and valid measurement of possible intervention effects and to exclude potential influences of different test items, the identical test items used in the pre-test were used in the post-test. Since the elapsed time between the pre- and the post-test amounted to a full year, learning or practice effects or memory biases were assumed to be minimal to non-existent. The research participants were again sent all necessary preparatory information in advance via e-mail. Because of the COVID-19 pandemic, the post-test fell into the first government-issued nation-wide lockdown (March – May 2020). Consequentially, students had to complete the post-test from home. As the test was administered via Moodle, I could ensure that the test was only accessible for a pre-defined limited amount of time. This way I could warrant that all participants completed the post-test at the same time under the same conditions (as far as this was controllable). Since the supervision of the test was not possible, all students were asked to fill in and sign a form to certify that they had not used any undue additional material (e.g., dictionaries, notes, etc.), and that they strictly adhered to the requirements throughout.

4.4.6. Data Processing

In preparation for the rating of the collected speech productions, the data needed to be processed. Since the aim of the pre- and post-test was to assess pre-service teachers' oral profession-related language ability based on their *performance*, conducting linguistic analyses

and thus transcribing the language productions was not necessary. Indeed, the attention of a performance test like the present lies on the overall test performance, i.e. the task completion in context, rather than on discrete linguistic aspects removed from the context (McNamara, 1996). Thus, I did not transcribe the data for several reasons. First, the raters' perception of the participants' performance, and consequently the evaluations thereof, could be skewed because a legible transcript could likely draw the raters' attention away from the actual, holistic task completion towards construct irrelevant factors. Second, transcripts would likely implicitly cause raters to map characteristics of written L2 productions onto oral language productions. Consequently, oral language productions could wrongfully be sanctioned for characteristics that are considered an "error" in written language, which are however "correct" features of oral language (see chapter 2.5.2.3). In speaking performance assessment in particular, the validity and utility of using controls of grammar or grammatical accuracy as an indicator of speaking proficiency is highly questioned (Luoma, 2009; Magnan, 1988). Indeed, when assessing oral language proficiency, communicative aspects are more meaningful and conclusive than features of sentence level grammar (Luoma, 2009). The focus of raters' judgements should therefore be on the communicative nature of the oral language performance, approaching complex linguistic behaviour and phenomena through a communicative rather than structuralist view of language (ibid.). Third, the assessment situation should be of high ecological validity. This aspect was also a requirement of the evaluation process. Judging a learner's oral language proficiency based on information perceived through the auditory channel rather than through reading a transcript increases the present test's ecological validity. Finally, I fully anonymised and randomised all pre- and post-test data so that associating single productions with t0 or t1 became impossible. I conducted these steps using MS Excel.

4.5. Scoring Test Performances: Rating

Competence-oriented performance tests such as the present serve the evaluation of complex human performance. To evaluate such performances, such tests necessarily involve judgements by expert raters based on a rating scale (McNamara, 1996). To ensure that the rating process generates valid judgements that allow for fair and reliable reporting of test scores, rigorous methods need to be applied (see chapter 2.5.4.4). One of these methods includes rater trainings that need to follow a systematic procedure tailored to the given test purpose, test construct and test format. The LSP test developed for this dissertation contains the PLLs and PLDs of the PRLCP and PRLC-R (see chapter 2.3.2 and 2.5.4.3) at its core. As the PRLC-R constitute the

basis for assessing the elicited performances, its descriptors thus constitute a dominant factor of the test purpose, test development and rating process (Cizek & Bunch, 2007). Indeed, they serve as one of the critical referents for the raters' judgements during the rating process (Cizek & Bunch, 2007). Considering the stakes of this test being relatively low, and with respect to both the considerations stated in chapter 2.5.5 as well as the test purpose and dominance of the PLDs in this research context, an appropriate method to set the standard (see chapter 2.5.5) and train raters constitutes Alderson, Clapham and Wall's (1995) method. The following section describes the rater training and rating processes employed in the present study. The chosen rater training method is adapted from and in line with the recommendations for standard setting and rater training procedures for constructed-response performance tests (see chapter 2.5.5).

4.5.1. Preparing the Rater Training

The adequate preparation of a systematic rater training involves considering a range of aspects. One of those is the stakeholders involved in and impacted by the test-development, the standard-setting and the rating process, as well as their strategic role. Another one is the selection of the experts that are to score the performances. Standard setting and rater training methods such as those introduced in chapter 2.5.5 require group decisions by experts, i.e. a standardising committee and trained raters. While the standardising committee is responsible for setting the standard, trained raters are responsible for marking the language productions. Considering the relatively low stakes and novelty of the present test, the standard setting and rater training were combined. The experts thus took on a double-role: they set the standard by identifying benchmarks, and undertook the rater training to score the pre- and post-test data. The next sections outline the steps taken to secure a sound standard-setting and rating process.

Selecting the Committee

As the process of setting (or more correctly: recommending) the standard involves group decisions, the composition of the committee and the training they receive is crucial to its success. Indeed, the committee members are deemed a key source of validity, credibility and variability of standard-setting results (Cizek & Bunch, 2007; Knoch & Macqueen, 2020). A first crucial step is therefore to select the committee to ensure that the members represent the purpose of the test (Knoch & Macqueen, 2020). As this standard-setting process builds on an LSP test, the involvement of domain experts is considered best practice (Knoch & Macqueen, 2020). The inclusion of domain experts is insofar valuable as they are the most likely to have a

sense of the minimum standard required to cope with the language demands of the domain (Knoch & Macqueen, 2020). They tend to base their judgements on the authentic professional context of the candidates, the social uses of language and the role of language in L2 teaching. This contributes to both preventing potential tensions between what a test measures and what is considered important in L2 teacher education – aspects that in turn contribute to achieving validity (Manias & McNamara, 2016). It is further recommended to include external committee members and stakeholders who represent different perspectives (AERA/APA/NCME, 1999). For the context of this dissertation, four committee members (R1, R2, R3 and R4) were selected based on to the following requirements (the information in the brackets showcase the committee members who meet the respective criteria):

- Experience in producing syllabus (R1, R2, R3, R4) and test specifications (R3);
- Experience in assessing productive skills against criteria (R1, R2, R3 and R4);
- Experience in developing language tests and writing items (R1 has experience in assessing language tests, R3, R4);
- Experience in coordinating groups of educators (R1, R3 and R4);
- Expertise in SLA and English Linguistics (R1, R2, R3 and R4);
- Experience in English language teaching (R1, R2 and R3) and English language teacher education at secondary school level (R1, R2, R3 and R4)

Subsequent to nominating the committee members, I contacted them individually, introduced them to the project and invited them to participate in the present research project. All members agreed to participate immediately. The selected group consisted of four female experts (average age at the time of setting the standard: 32.5) who all work as lecturers and researchers in the fields of Applied Linguistics and English language teaching at the Department of Language Didactics (Institut Fachdidaktik Sprachen IFDS) at the PHSG. All committee members had acquired at least a Master's level qualification in English Language and Linguistics (R1, R2), English Literature (R3, R4), Educational Psychology (R1), or Pedagogy (R2). While one committee member (R1) was a PhD graduate in English Language and Cognitive Linguistics and a post-doctoral student in the respective field, two members were doctoral students in English Linguistics (R2) and Foreign Language Teaching and Learning (R3). The final member (R4) was the head of the English subject group at the PHSG and has extensive experience in English language teacher education at the target level of lower secondary school. R2 and R3

obtained an additional teaching qualification to teach English at high school level. R1, R3 and R4 had completed some or all requirements to achieve their Tertiary Education Teaching Qualification. The average teaching experience of the committee at the time of setting the standard was 5.25 years and the experience with examinations averaged 4 years. R1 and R4 had experience in assessing productive skills in relation to defined criteria in both higher stakes (matura examinations, mid-term and end-of-term examinations in Teacher education programmes) and lower stakes contexts (formative evaluations in a range of educational contexts). R1, R2 and R3 had experience in undergoing comprehensive and large-scale rater trainings to evaluate productive language skills in high stakes language tests. R1, R2 and R3 can be considered domain experts, although due to a limited amount of experience in the field it can be argued that the domain expertise of the committee was restricted. However, the overall required expertise of the group was extensive and the individual members complemented each other with their specific experience and knowledge to a satisfactory level. Although it is recommended to include external committee members external to represent different viewpoints, and although R1 teaches at an external higher education institution, there were no entirely external members on the panel.

Familiarising the Committee

After establishing the committee, appropriate training was needed to achieve the training's ultimate goal of aligning the judgements so that they become reproducible and fair. The Standards for Educational and Psychological Testing emphasise the importance of the committee understanding of what they are to do: "The process must be such that well-qualified judges can apply their knowledge and experience to reach meaningful and relevant judgements that accurately reflect their understandings and intentions" (AERA/APA/NCME, 1999, p. 54). Appropriate standardising committee and rater training includes introducing the committee to the purpose and the test format and the decision-making levels that need to be set in advance of setting the standard (Knoch & Macqueen, 2020). The training also needs to provide ample opportunity to "explicitly discuss and adopt a position on the issue in advance of standard setting" (Cizek & Bunch, 2007), which in turn needs to strongly relate to "the purpose of the examination and the classifications that are to be made based on the cut score(s)" (Cizek & Bunch, 2007). As previously introduced, Alderson, Clapham and Wall's (1995) approach to standard-setting and rater training built the framework for the local implementation of the present rating process. The recommended procedure was slightly adapted and recommendations of the Council of Europe (2001) were incorporated to suit the purpose of this language-testing

situation. As stated above, this method is characterised through the dominant role of PLDs during the standard setting and rater training. In addition, a chief examiner (CE) takes a leading role in the process. The CE is responsible for the rationale to be followed and the raters selected, leads the rater training and ensures that consensus on the benchmarks is reached by all committee members (Alderson, Clapham & Wall, 1995). The CE forms part of the rating committee and is involved with the subsequent rating process. In this local application, I took on the role of the CE. In preparation for the rater training, I rated and analysed a number of spoken productions to become familiar with the test performances and the problems the test takers experienced in completing the tasks. This step also served to set initial benchmarks as a rough guideline for the process to follow. With the rating scale in mind, I extracted 21 spoken productions (3 productions per test task) which represented "adequate", "satisfactory" and "inadequate" performances. All marked language productions were divided into two batches (batch A and batch B). Subsequently, I sent the following necessary informational documents to the rating committee:

- an introduction to the background and underlying competence model of the test and test construct in question (PRLCP),
- the test specifications and detailed test task explanations,
- the PRLC-R (PLDs with clear explanations),
- a comprehensive rating manual (see appendix D),
- a statement of purpose for setting standards and training raters,
- the previously selected sample test answers taken from batch A,
- an individual assessment sheet to record rating results, and
- comprehensive instructions for the familiarisation task.

It is widely considered effective for the members to have experience with the assessment on which they will make judgements, even though from a validity evidence perspective it would seem appropriate not to have any experience with the assessment at all (Cizek & Bunch, 2007). Therefore, I gave the committee members access to the entire test for self-administration in preparation to the meeting. This initial stage is referred to as "familiarisation" and allows the committee members to become acquainted with the context and the purpose of the training, and the demands placed on test takers. At this stage, the committee members were asked to carefully

study all materials provided, self-administer the pre- and post-test, record any questions and comments with regard to any of the provided documents, and use the PRLC-R to mark all spoken productions of batch A to set and record benchmarks individually on the respective assessment sheet. They were also asked to take notes on why they awarded their scores on each of the seven rating criteria. In order to ensure that all participants had the same exposure to the rater training method and received this information under uniform conditions, no information regarding the rater training method was transmitted to the committee prior to the meeting (Cizek & Bunch, 2007). Once the committee members completed the familiarisation task, an entire day was initially set aside for the rater training to proceed shortly before the official marking period began.

4.5.2. Conducting the Rater Training

Standard-setting meetings and rater trainings must follow a set structure. The present procedure (cf. Alderson, Clapham & Wall, 1995) contained 3 distinct stages. At stage 1, I (i.e. the CE) held an introductory presentation to reiterate the purpose of the standard setting and summarise the contents of the informational documents committee members had received in advance. A clear outline of the agenda for the day and a description of what participants were going to be asked to perform was then presented. Stage 2 commenced by clarifying any questions regarding the test (items, construct and purpose), the PRLC-R and the rating manual, and by adjusting the latter two documents if needed. This additional discussion and refinement of the PRLC-R and PLDs served to make them more user-friendly and to facilitate understanding. The subsequent step was to compare the marked language productions of batch A and discuss differences of opinion. Ample room for open discussion was necessary and this step was by far the most time-consuming of the entire process. The aim of this central stage was to reach a *consensus mark* for each response of batch A in terms of assigning them to a performance level. This consensus mark would then be included in the rating manual and serve as a benchmark for the subsequent rating process. Benchmarks are considered particularly useful tools for aiding consensus-building before, and maintaining consensus during the rating process (Arras, 2011). In further discussions, the committee members agreed on how to proceed when marking problem productions. I recorded these decisions and included them in the rating manual. Due to the discussions being extensive and the process of reaching consensus among all committee members taking up a lot of time, this stage could not be completed on the day initially planned

for the entire rater training. Hence, an additional meeting[22] was scheduled one week later, which served to complete the discussion of the batch A ratings (finalisation of stage 2) and to complete stage 3. During this stage, the members individually marked some of the language productions from batch B for further practice. The purpose of this stage was to solidify members' understanding of the application of the rating instruments and to foster agreement through another discussion. Once consensus on all production marks was reached, I collected all members' reasons for each of their decisions, tallied the results and recorded them in the rating manual in form of step-by-step guidelines to ensure that they were available as benchmarks. To complete stage 3, I explained the procedures for recording marks when rating. After the rater training, the PRLC-R was finalised by incorporating the agreed changes. These included minor adjustments to the wording of some descriptors and the addition of a precise descriptor on level *0* for each criterion. Prior to the training, level *0* contained the generic descriptor "Ausführung (noch) nicht wie auf level *1*" (task execution does not (yet) reach level *1*) and was visually underrepresented. I broadened the column to equal its width to the performance levels *1, 2* and *3.* This served to give level *0* more visual presence because the raters had reported that they had previously overlooked it. In addition, the subdivided performance levels were simplified so that instead of each level being subdivided into two, only level 2 offered a subdivision: *0, 1, 2, 2** and *3* were now possible to select:

Ausführungsniveau Qualität der Aufgabenausführung	**0**	**1**	**2**		**3**	**Bemerkungen**
Wortschatz: Wortwahl Sich im gegebenen Kontext mit inhaltlich passender Wortwahl ausdrücken	Ihre/seine Wortwahl ist im gegebenen Kontext **durchgehend inhaltlich unpassend.**	Ihre/seine Wortwahl ist im gegebenen Kontext **wiederholt inhaltlich unpassend**.	Ihre/seine Wortwahl ist im gegebenen Kontext inhaltlich **grundsätzlich passend**.		Ihre/seine Wortwahl ist im gegebenen Kontext inhaltlich **differenziert** und **treffend**.	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	
Sprachliche Korrektheit Sich sprachlich korrekt ausdrücken (Grammatik)	Sie/er macht **so häufig** grammatische Fehler, dass **durchgehend unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **häufig** grammatische Fehler, wobei **teilweise unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **manchmal** grammatische Fehler, wobei **grundsätzlich klar** ist, was sie/er ausdrücken möchte.		Sie/er macht nur **sehr selten** oder **gar nie** grammatische Fehler, die auffallen.	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	

Figure 18 : Excerpt finalised version of the PRLC-R for the rating period

4.5.3. Rating the Test Performances

Immediately following the rater training and prior to the beginning of the rating process, I administered a Google Forms questionnaire. This form of summative evaluation served to

[22] Day two of the standard setting and rater training meeting fell into the onset of COVID19 restrictions imposed by the Swiss Government in April 2020. Hence, as opposed to day one, this second part of the meeting was conducted online via Skype.

obtain the committee's overall reactions to various aspects of the meeting and to measure their confidence and support for the final group recommendations. This step is considered crucial to supporting the validity of the procedure and provides retrospective insights into reasons for potential rater disagreements. Results showed a high degree of rater satisfaction and confidence with the effectiveness of the rater training conducted (3/3 selected level 5 as the highest level of satisfaction) and its outcome (1 member selected 4/5, 2 members awarded 5/5). Questions that remained unclear after the training mainly concerned rater biases and rater effects (e.g., favouring certain accents over others) and the difficulty of minimising subjectivity while rating. Raters mentioned that the detailed rating manual was helpful for recognising subjective perceptions and for mitigating those while rating. The raters indicated a high degree of familiarity with the rating scale (1 member selected 4/5, 2 members awarded 5/5) and reported experiencing a reasonable level of difficulty when judging the language productions (2 member selected 3/5, 1 member chose 4/5). Some of the difficulties experienced during the rater training were managing to maintain a clear distinction between the individual PLDs and between the overall performance levels, and to apply these distinctions rigorously and consistently. There was a reported awareness that the boundaries between the criteria and between the performance levels were not clear-cut. The criterion *addressee-specificity* proved to be the most difficult to fully grasp, as reported by 2 members. All members however reported a high level of satisfaction with the group discussions and indicated that the critical discourse among participants had been helpful for understanding the PLDs and performance level thresholds. Further questions arose with reference to the cognitive load. Issues such as structuring their own rating practice in a way to keep concentration high and the rating structured were mentioned. The degree of confidence with the PRLC-R was reportedly satisfactory (1 member selected 4/5, 2 members awarded 5/5) and all members were satisfied with the amount of time that had been dedicated to the group discussions, as well as with the final consensus reached.

In addition to a summative evaluation, it is considered good practice and essential to ensuring validity to provide (confidential) feedback to the committee members following their judgements about items, tasks and examinees during the standard-setting process (Cizek & Bunch, 2007). This step was conducted recurrently during the rater training and subsequently on a more low-threshold level during the rating process itself (e.g., by reviewing individual ratings and answering individual questions, or by commenting on raters' judgements and judgement explanations in the comments box). I coordinated the rating process by sending each rater a new assessment sheet for a new batch of around 20 language productions on a weekly

basis for the duration of 7 weeks. I made the audio files available to each individual rater in a shared folder on SWITCHdrive. The raters then sent the completed assessment sheets back to me each week, which I carefully reviewed, recorded, and – if applicable and necessary – provided brief feedback or individual answers to (see above). 40% of the data were rated double-blind. Halfway through the rating period, an online rater conference was held in order to discuss problems that had emerged up to that point. By analysing the present ratings, the components *vocabulary*, *pronunciation*, *coherence & cohesion* and *addressee-specificity* of the assessment scale tended to be rated with larger discrepancies between raters. One of the main difficulties was to differentiate between individual components: deciding, for example, whether a test taker's response including a false description of a grammar rule should be marked in *task achievement* (not meeting the task requirements), *accuracy* (inaccurate grammar knowledge) or *vocabulary* (choice of inaccurate or unsuitable vocabulary). These components thus needed to be further clarified. Moreover, a number of questions arose that were task-specific and needed further discussion to reach consensus for the ongoing rating process. A number of task responses that revealed differing ratings were revisited and the decisions for each rating were discussed. Finally, consensus was reached and specifications were added to the rating manual for future reference. The ratings conducted prior to the interim rater conference were not reassessed and readjusted after the conference for research-economic reasons. While this is certainly a limitation to the overall rating process, the scope of the project did not allow for this additional step. After the rater training and marking period, I assessed the degree of agreement amongst the participants by statistically analysing the ratings (see chapter 5.1.1).

4.6. Summary Research Methodology Main-Study

The aim of the main-study was to empirically investigate the implementation and application of the PRLCP and PRLC-R in a relevant L2 education context in order to determine their affordances and challenges with reference to the development of oral, profession-related language competences in teacher education. I devised a quasi-experimental pre-post design, which encompassed an experimental-control intervention. An entire PHSG cohort participated in the main-study while completing their BA E-Portfolio part B, a compulsory multi-draft assignment that involves students to iteratively work on their oral, profession-related language competences over the course of one year. For this purpose, I developed a near-authentic, competence-oriented performance test, which was administered online before and after the treatment. To score the test responses, I conducted a rigorous rater training. Subsequently, 40%

of the language performances were rated double-blind, which then allowed for a MFRA to control for rater effects and interaction effects and to analyse the pre- and post-test data.

5 Data Analyses and Results Main-Study

L2 teaching and learning research is characterised through its multifaceted and highly complex nature (Grum & Zydatiss, 2016). As many possible influential and confounding variables are prevalent, the correlations and connections between them are not always recognisable and many variables are not directly measurable (Grum & Zydatiss, 2016, p. 322). Empirical investigations in this field are therefore confronted with having to take into account and control the multitude of variables that reciprocally influence one another (Grum & Zydatiss, 2016). This chapter contains a description of the ways in which I attempted to do so and outlines the statistical tools and methods I employed to analyse the data obtained in the main-study. All statistical analyses were conducted using R Studio of the open-source software R (RStudio version 3.6.0., 2019, www.r-project.org) with the extensive guidance of a language testing expert. The first section (chapter 5.1) presents the analyses of the main-study and contains three parts. The first part (5.1.1) investigates the interrater reliability of the overall ratings. The second part (5.1.2) presents further investigations into differential rater functioning, with a focus on rater severity, within-rater consistency, gender bias, and interaction analyses between raters and test takers, rating criteria, and test tasks. In the third part (5.1.3), I present the analyses of pre-service English teachers' oral profession-related language competences based on their test task responses of the pre- and post-test. These analyses serve to uncover areas of competence in which the treatment of the main-study affected the research participants' oral profession-related language performances over time, and to determine the extent to which possible effects are observable. The results of all the analyses above and the answers to RQ #1, RQ #2.1, RQ #2.2 and RQ #2.3 are outlined in the subsequent section (chapter 5.2).

5.1. Analyses Main-Study

This section contains all analyses conducted in the main-study. The first set of analyses relates to the expert ratings where the focus centred on computing reliability coefficients to determine

the interrater reliability of all ratings conducted. The second set concerns the rater main effects as well as 2- and 3-way interactions between raters, examinees, rating criteria, and tasks. The third set of analyses relates to the estimation of individual test taker proficiencies, rater severity and leniency, and criteria and task difficulty, and scale category difficulties. The final set of analyses concerns participants' pre- and post-test performances and the comparison of within- and between-group results from t0 to t1 to identify whether and to what extent the treatment affected the participants' oral teacher language competences. The present analyses are based on the ratings of overall 33 participants' pre- and post-test responses. One student dropped out throughout the intervention study, leaving 34 students at t0 and 33 students at t1. The data from the dropout participant were excluded from the analyses. Out of 445 audio-file test responses, 30 responses were unusable because of bad sound quality, leaving 415 usable files for the overall data analysis.

5.1.1. Interrater Reliability

To identify the usability of the data for the subsequent pre-post-test analysis, the reliability of the obtained ratings such as their stability, reproducibility, and accuracy needed to be determined (see chapter 2.5.4.4). The reproducibility of the ratings as a type of reliability is considered one of the strongest and most feasible types to test (Krippendorff, 2004a). One way of doing so is by calculating the interrater reliability (IRR) between and across all independent raters who evaluated the performances (R1, R2, R3 and R4). There are a multitude of statistical tests that can be used to measure IRR (Eckes, 2005; Hayes & Krippendorff, 2007). Because different IRR coefficients respond to different properties of the data and therefore reflect different information, it is advisable to report on a variety of different coefficients to gain a more comprehensive understanding of the data structure (Wirtz & Caspar, 2002). Because the present data are ordinal data, it is advisable to compute Krippendorff's Alpha (α) (Wirtz & Caspar, 2002). Krippendorff's α calculates rater *dis*agreement and is flexible in that it accounts for agreement by chance and for more than two raters. It is compatible with ordinal, interval, ratio or nominal data (Hayes & Krippendorff, 2007; Krippendorff, 1970, 2004a, 2004b; O'Connor & Joffe, 2020). Krippendorff's α can also be used regardless of the sample size and the absence or presence of data (Hayes & Krippendorff, 2007). Cohens κ (Cohen, 1960) constitutes a possible complementing alternative, however there are some limitations. Cohen's Kappa (κ) calculates the proportion of exact agreement on a scale of $-1 \leq \kappa \leq 1$ and takes into account the element of chance. Cohen's κ is however restricted to merely reporting on percent

agreement[23] of two raters and nominal data, thus it is not applicable for the present analyses (Eckes, 2011; Fleiss et al., 2003; Hayes & Krippendorff, 2007; Wirtz & Caspar, 2002). Another option offers Pearson's product-moment correlation[24]. This test measures the strength and direction of association that exists between two variables (here: judgements of two raters) measured on at least an interval scale ($-1 \leq r \leq 1$) provided that there is a linear relationship between the two variables in question. Negative values generally indicate inverse judgements where the interrater reliability is equated to 0 (Eckes, 2011; Wirtz & Caspar, 2002). Pearson's correlation *r*, however, does not allow for correcting the probability that a certain amount of agreement occurs by chance. Opposed to Cohen's κ and Pearson's correlation *r*, Krippendorff's α proved to be the most suitable coefficient for the present and available data due to its flexibility and was thus chosen for conducting the IRR computations. To calculate Krippendorff's α, all double ratings (40% of the entire data pool: 198 audio files rated against 7 rating criteria, resulting in 1386 single data points) were recoded from 0, 1, 2, 2* and 3 to 0, 1, 2, 3 and 4. In R, the function kripp.alpha() does not compute a confidence interval. In order to consider the confidence interval when computing α, the bootstrap method can be applied by using the function kripp.boot(). The bootstrap resampling method of subjects is a function that can be used to obtain valid standard errors (Berk et al., 2014). By bootstrapping the distribution of α from the given reliability data, one can thus avoid merely assuming approximations (Hayes & Krippendorff, 2007). While bootstrapping allows for standard errors, it is important to keep in mind that the values obtained are still a mere estimation, and that estimations always contain errors. In a next step, Krippendorff's α was calculated across all rating criteria for rater pairs. Finally, the agreement between all raters was calculated per rating criterion. For this step, the ratings were recoded once more by collapsing level 2 and 3 to one level. The corresponding results are presented in chapter 5.2.1.

5.1.2. Bias and Interaction Analyses

Interrater reliability measures are helpful to determine the agreement with which raters evaluate performance samples. They are important in terms of determining the validity, objectivity and reliability of a test and of ratings. However, the statistics do not allow to measure *rater effects* (raters' unique perceptions and rater bias) and *interaction effects* (e.g., the influence of scale

[23] The percent agreement is "the proportion of units with matching descriptions on which two observers agree" (Hayes & Krippendorff, 2007, p. 80).
[24] Pearson's correlation; ρ = when measured in the population and *r* = when measured in a sample

characteristics and test task variation on rating behaviour) that typically lead to systematic rater variability (Lunz & Stahl, 1990) (see chapter 2.5.4.4). Previous studies have shown that a large proportion of variance in ratings can be ascribed to rater effects (Hoyt & Kerns, 1999), and that a *Multi-faceted Rasch analysis* (MFRA) enables modelling rater variation and allows to compensate for these sources of variations (McNamara, 1996). Thus, an MFRA can be used to conduct bias and interaction analyses that can help "to identify unusual interaction patterns among facet elements, particularly those patterns that point to consistent deviations from what is expected on the basis of the model" (Eckes, 2005, p. 203). The following analyses align with Eckes' (2005) approach to examining commonly identified rater effects in TestDaF[25] writing and speaking performance assessments. They build on an MFRA and aim to investigate common main effects and biases as identified in the literature (cf. Eckes, 2005; Hoyt & Kerns, 1999) through the following questions (cited and adapted from Eckes, 2005):

RQ #2.1 Do the raters differ in the severity or leniency with which they rate the test takers' performances?

a) Does each rater maintain a uniform level of severity, or do particular raters score more harshly or leniently than expected?

RQ #2.2 Do the raters maintain a uniform level of severity or leniency across criteria and across tasks?

b) Do ratings on one criterion follow a pattern that is markedly different from ratings on the others?

RQ #2.3 Do the raters show evidence of differential rater functioning related to test takers' gender; that is, do they maintain a uniform level of severity or leniency across male and female test takers?

To answer questions #2.1-#2.3, a partial credit model (PCM) was implemented. The PCM specifies that each criterion of the PRLC-R has its own rating scale structure, thus allowing the PRLC-R to vary with reference to the different criteria. Item (here: PRLC-R criterion), task and rater were modeled as facets. While in the IRR analyses the performance levels 2 and 2* were collapsed into one, this step was not conducted for the MFRA because they both contained enough observations for computations. Instead, category 0 and 1 were collapsed into one due to the small amount of observations in each, and the data was recoded from 0, 1, 2, 3 and 4 (see

[25] TestDaF is a renowned large-scale, high-stakes international certificate for German as a foreign language. See https://www.testdaf.de/de/ for more information.

chapter 5.1.1) to performance levels 0, 1, 2, and 3. Collapsing ordinal categories is common (Healey, 2012) and reasonable as long as that data structure is not altered (Kateri, 2014). After estimating the parameters of the model, an examination of the standardised residuals (i.e. standardised differences between the model-based expected ratings and observed ratings) was conducted to identify possible bias and interaction effects (Eckes, 2005, p. 203). In bias analyses, (1) individual rater severity or leniency across all tasks and all criteria, (2) overall rater severity or leniency per individual criterion, and (3) overall rater severity or leniency per individual task were examined. In two-way interaction analyses, the interactions between Rater × Criterion, and Rater × Task were examined to test if the combination of a specific rater and specific criterion or task led to (inconsistent) rater severity or leniency (Eckes, 2005). In addition, a gender bias statistic (Z statistic) was computed to test whether raters exercised differential severity or leniency depending on whether they judged a female's or a male's language production, or whether the level of severity or leniency remained stable across gender groups. The test takers were treated as random effect throughout. The results of these analyses are presented in chapter 5.2.2.

5.1.3. Pre-Post Analyses

To compare the pre- and post-test results to identify possible treatment effects, a competence comparison was conducted by means of the partial credits model (PCM) without any further interactions and no adjustment for multiple testing. Competence differences were investigated at t0, at t1 and from t0 to t1 within and between the experimental group (E), control group 1 (C1) and control group 0 (C0). Instead of a two-dimensional model, a one-dimensional model was computed using person IDs (PID) treating the research participants at t1 as different persons to the research participants at t0. The data from the C1 at t0 were used as the overall reference point to determine any treatment effects. To make the group comparisons of the interaction model more transparent, estimated means by means of weighted likelihood estimation (WLE, no adjustment for multiple testing, cf. Warm, 1989) were calculated. First, an overall within and between-group competence comparison was calculated at and between t0 and t1 across the test takers' overall pre- and post-test performance. Second, the same calculation was conducted across all treatment groups, but per individual rating criterion (i.e. competence dimension) at and between t0 and t1 to identify any competence development in specific areas of performance. Finally, competence comparisons at and between t0 and t1 were conducted for each treatment

group individually and per individual rating criterion. The results of these analyses are presented in chapter 5.2.3.

5.2. Results Main-Study

In the following section, I present the results of the data analyses described above. I begin with outlining the results of the interrater reliability calculations, proceed with the findings of the interaction analyses, and conclude with the results obtained from examining potential treatment effects within and between groups between t0 and t1.

5.2.1. Results Interrater Reliability

This subsection presents the results of all IRR calculations. Overall, the observed mean α level for four raters (R1, R2, R3 and R4; 5000 iterations) is α = 0.338 at a confidence interval of 0.257 and 0.415. This means that there is a 95% chance that the observed mean α value of the population lies between 0.257 und 0.415. When it comes to interpreting the computed values, Landis and Koch's (1977) recommendation is often considered as the convention (O'Connor & Joffe, 2020). On a scale from 0 to 1, they recommend that an index less than 0 indicates no, between 0 and 0.20 slight, 0.21 and 0.40 fair, 0.41 and 0.60 moderate, 0.61 and 0.80 substantial, and 0.81 and 1 near-perfect agreement. The absence of agreement means that there is no statistical relation between the units of analysis and how they were identified, coded, or described (Hayes & Krippendorff, 2007). A more conservative interpretation is offered by Koo and Li (2016) who state that IRR values below 0.50 indicate poor, between 0.50 and 0.75 moderate, between 0.75 and 0.90 good and above 0.90 excellent agreement. I adopt Landis and Koch's (1977) interpretation for the following interpretation. In any case, whether one orientates the interpretation of the values according to Landis and Koch (1977) or Koo and Li (2016), the computed mean α value of 0.338 indicates fair, or poor agreement, respectively. To further investigate these results, a contingency table was created to display the frequency of assigned ratings across all double ratings by any given two raters. The aim of a contingency table is to visualise to what extent the first and second rating of any double-rated performance correspond to one another. Ideally, all cells outside of the green diagonal cells contain 0 matches, which would indicate a perfect agreement (P_o, proportion of exact corresponding judgements, cf. Eckes, 2011) between any given rater in any double rating. The following table

highlights two main findings: 1) there is a low amount of observed perfect agreement between raters, and 2) there is an obvious ceiling effect:

	0	1	2	3	4
0	0	1	0	1	0
1	0	16	33	15	15
2	1	14	52	79	53
3	0	16	76	141	117
4	0	18	60	176	360

Table 9 : Contingency table of perfect agreement between two given raters

Said observed ceiling effect becomes apparent in the following box plots that display the frequency of assigned ratings on the level of the individual rater:

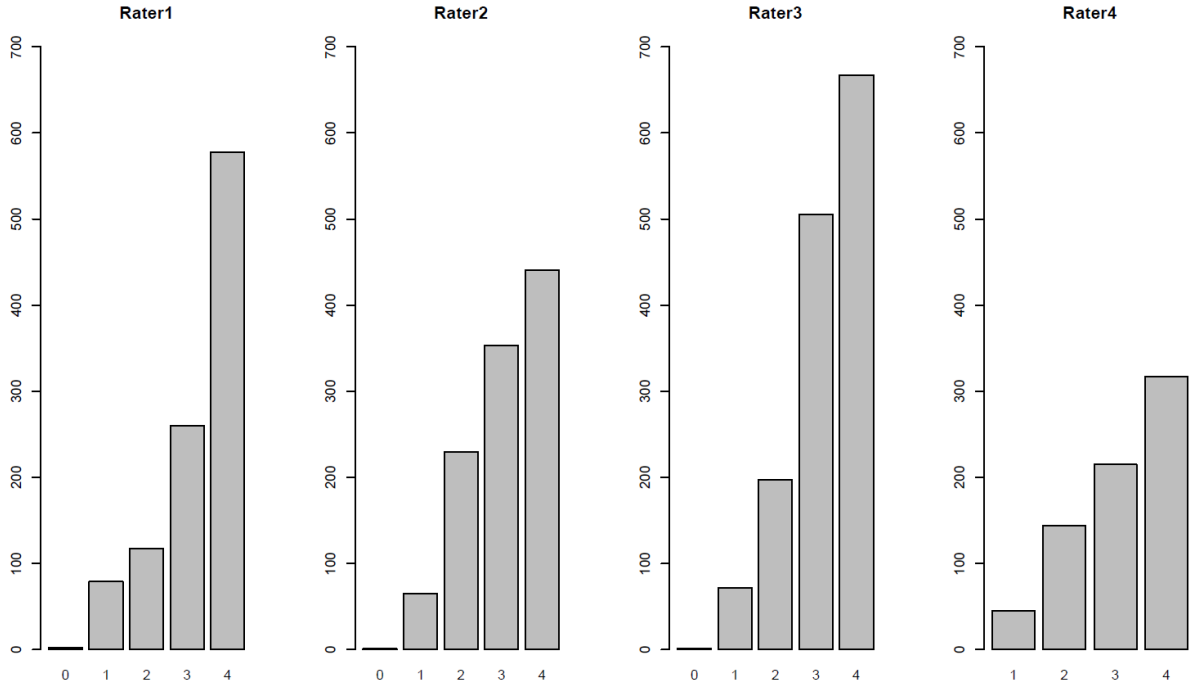


Figure 19 : Frequency of levels assigned to productions by individual raters

Note here that not all raters judged equal amounts of language productions. R4 judged significantly less, while R1, R2 and R3 evaluated a comparable amount. Thus, the above plots cannot be compared with reference to relative frequency. In any case, ceiling effects are not necessarily to be interpreted as a negative result with reference to rating behaviour (it could for

example have been caused through generally high competence levels of test takers at both t0 and t1). However, since the present competence-oriented performance test's purpose was to uncover the variance among test takers' competences (as opposed to, for instance, an achievement test, see chapter 2.5.1), the ceiling effect could also be an indication of rating scale or rating behavior problems (see chapter 6.2). In further analyses, the subsequent calculations of Krippendorff's α across all rating criteria for rater pairs reveal the following results:

Rater Pair	*Krippendorff's α*
rater 1 & rater 2	0.257
rater 1 & rater 3	0.262
rater 1 & rater 4	0.263
rater 2 & rater 3	0.416
rater 2 & rater 4	0.430
rater 3 & rater 4	0.305

Table 10 : Krippendorff's α for rater pairs

The results indicate the rater pairs R2 and R4 show the highest agreement, closely followed by R2 and R3. In addition, the results show that R1 and R2 rated most differently from one another overall. These values thus suggest that on average, both R1 and R4 tended to diverge more in their ratings from the overall rater group than R2 and R3. Finally, the Krippendorff's α values to indicate the agreement across all raters per rating criterion reveal the following (illustrated from lowest levels of agreement to the highest):

Criterion	*Krippendorff's α*
Addressee-specificity	0.0806
Vocabulary	0.15
Pronunciation	0.298
Coherence and cohesion	0.308
Task completion	0.359
Accuracy	0.399
Fluency	0.501

Table 11 : Krippendorff's α across all raters per rating criterion

When interpreting these values against Landis and Koch's (1977) recommendations (0 = no, 0-0.20 = slight, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial, 0.81-1 = near-perfect agreement), only the agreement on *fluency* can be considered moderate. All other criteria fall below the moderate rate, indicating that raters did not judge reliably. However, even though the reliability values indicate poor reliability overall, between individual rater pairs, and across all individual evaluation criteria, it is important to treat the interpretation of computed

reliability coefficients with caution. Just as high agreement between raters alone does not exclude the possibility that all raters subconsciously rated with the same error (Eckes, 2011), low agreement as evident in the present analyses merely represents a number without taking into account rater variability caused by rater or interaction effects. Thus, this poor result does not immediately mean that the data is entirely unusable. First, additional analyses are necessary. To further investigate what lies behind the observed α values, the next section reports on more detailed analyses undertaken to shed light on potential sources of variability.

5.2.2. Results Bias and Interaction Analyses

This section presents the results of the bias and interaction analyses conducted by means of an MFRA (partial credit model) to identify differential rater behaviour and rater variability with reference to rating criteria and test tasks. In a Rasch context, (mean-square) fit statistics indicate how accurately or predictably a set of data fit the model, i.e. the amount of randomness or distortion of the measurement system (Linacre, 2002). 1.0 is their expected value, and fit statistics can range from 0 to infinity (Linacre, 2002; Myford & Wolfe, 2003). According to Linacre (2002), mean-square values near 1.0 indicate little distortion of the measurement system, values below 1.0 indicate that observations are too predictable (redundancy, i.e. the data *overfit* the model), and values above 1.0 indicate that the observations are unpredictable (unmodelled noise, i.e. the data *underfit* the model). *Global model fit* statistics allow for the assessment of the overall data-model fit, i.e. the extent of unexpected ratings across all data given the assumptions of the model (Eckes, 2005). In addition, mean-square item fit statistics can be reported to indicate data–model fit (*rater fit*) for each rater: *rater infit* and *rater outfit*. While rater infit is sensitive to an accumulation of unexpected ratings (inlier-sensitive fit), rater outfit is sensitive to individual unexpected ratings (outlier-sensitive fit, cf. Linacre, 2002). The lower the infit values, the more conservatively the raters employed the criteria. In other words, the less variable the ratings were than expected based on the model, the less the raters took advantage of the entirety of the scale and the more they showed a tendency to the middle. As outlined with the global model fit statistics, such data also tend to *overfit* the model. In contrast, fit values above 1.0 indicate more variation than expected in their ratings. Such data tend to *misfit* (or *underfit*) the model. Linacre (2002) suggests using 0.50 as a lower, and 1.50 as an upper control limit (Eckes, 2005). *Rater fit* in the present fit statistics thus indicates the degree to which a rater displays overall unexpected rating behaviour, i.e. the extent to which raters are systematically internally consistent and apply the PRLC-R appropriately after individual rater

severity and leniency is controlled for (Eckes, 2005). I will now proceed to report the results of the fit statistics and bias and interaction analyses in the sections below.

Global Model Fit

It is possible to assess an overall data–model fit by investigating the unexpected ratings given the assumptions of the model (Eckes, 2005). A model fit is considered satisfactory "when about 5% or less of (absolute) standardised residuals are equal or greater than 2, and about 1% or less of (absolute) standardised residuals are equal or greater than 3" (Eckes, 2005, p. 204; cf. Linacre, 2004). In the present case, out of 4289 valid responses, 319 responses (i.e. 7.4%) are associated with (absolute) standardised residuals equal or greater than 2. Furthermore, 182 responses (i.e. 4.2%) are associated with (absolute) standardised residuals equal or greater than 3. This means that in 7.4% (instead of 5% or less) and 4.1% (instead of 1% or less) the estimation of the deviation is significant, respectively. Thus, these findings indicate a not quite satisfactory overall model fit. In other words, a reasonable amount of rater responses was unexpected given the assumptions of the model and the raters displayed differential interactions and deviation from the model.

Calibrations of Test Takers, Raters, Criteria, and Tasks

The below Wright Maps (Figure 20 and Figure 21) display the mean values of the 50% Thurstone Thresholds (item thresholds) of the seven PRLC-R assessment criteria, i.e. the calibrations of the test takers, raters, and rating scale criteria as raters used the PRLC-R to score the participants' spoken task responses. Thurstone Thresholds are one possible way of computing rating scale category boundaries (Linacre, 1998, 2003; Thurstone, 1928). They indicate the position on a scale where there is a 50% chance that a rater allocates one of two adjacent performance levels (Eckes, 2011). Here, they indicate the breadth with which the raters interpreted and applied each performance level across task and test takers. Figure 20 displays the combined student distribution and 50% Thurstone Thresholds and shows that overall, the raters interpreted the criteria thresholds substantially differently across the PRLC-R criteria. For instance, the criterion *content* (con) was interpreted much more narrowly by the raters than *coherence & cohesion* (coh):

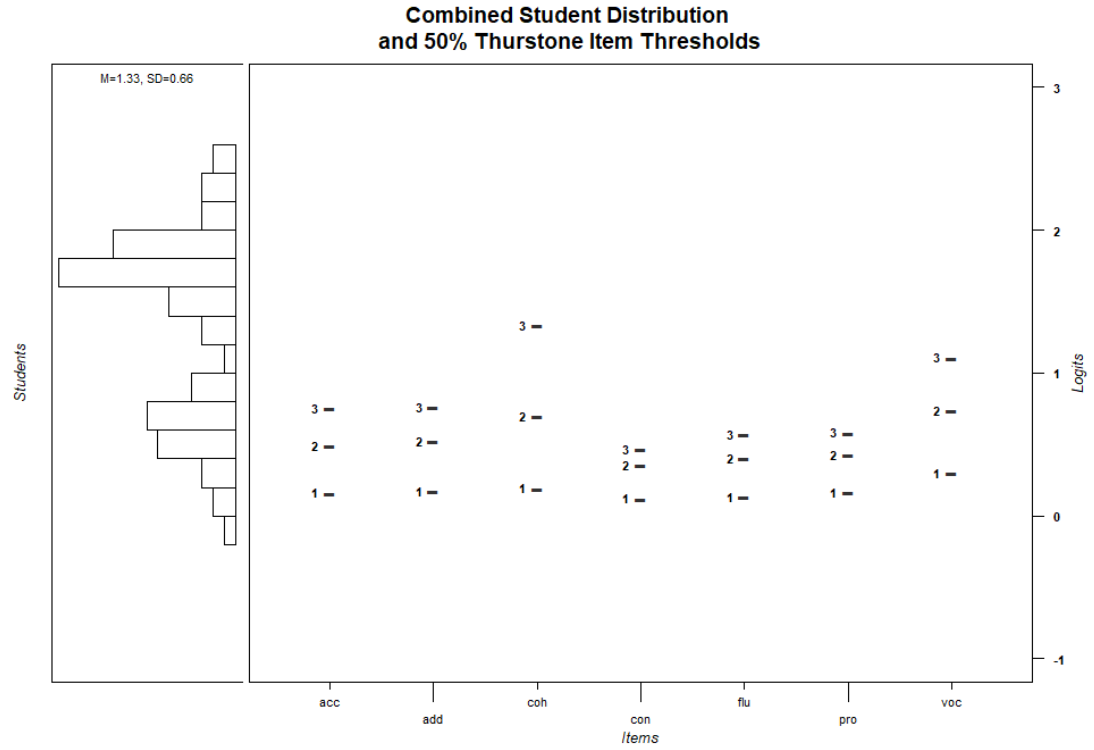


Figure 20 : Combined test taker distribution and item thresholds

For instance, the left-most thresholds in Figure 20 indicate that the probability that a test taker is allocated to performance level *2* for *accuracy* is equal (50%) to the probability that she or he is awarded a *1* or a *3*. Figure 21 below provides further insights into individual raters' interpretations and their application of the PRLC-R criteria. It indicates how each individual rater interpreted the performance level boundaries across each rating criterion, task and test taker. Reading the figure from left to right, the values indicate each rater's performance level boundary interpretation with reference to one particular criterion (e.g., *accuracy* judged by rater 1 = accr1, then by rather 2 = accr2, and so on). As can be seen, the variability across raters in their level of severity (i.e. leniency) was substantial. Two performance level boundary interpretations that particularly stand out are those in relation to *addressee-specificity* and *coherence & cohesion*. For instance, rater 2 interpreted the performance level boundaries of *coherence & cohesion* (cohr2) substantially different from rater 4 (cohr4). While rater 2 judged a test taker's performance against the criterion *coherence & cohesion* to be at performance level 1, rater 4 allocated that same test taker's performance to level 3 with reference to the same criterion, indicating significantly higher leniency in comparison to rater 2 (see also chapter 6.1):

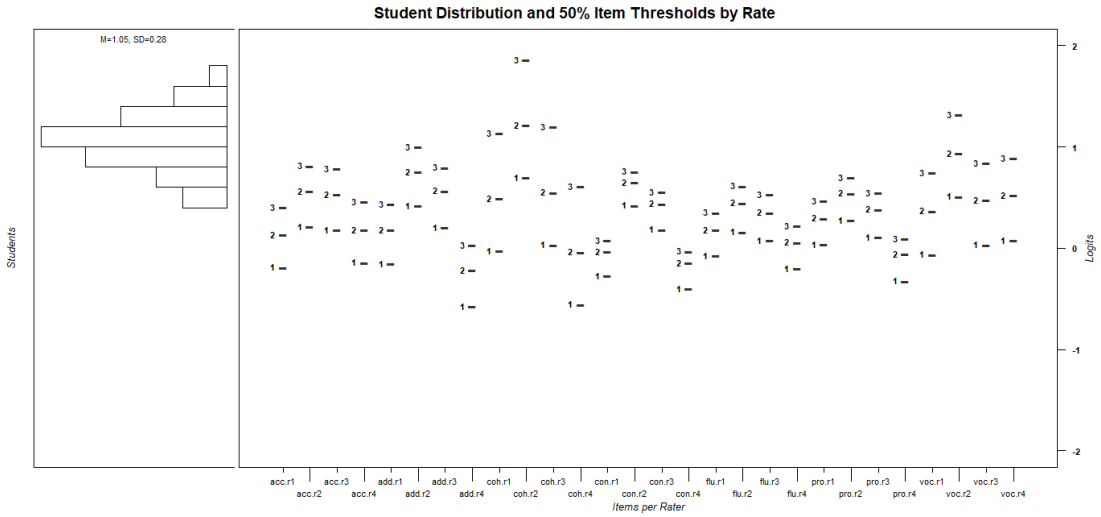


Figure 21 : Test taker distribution and item thresholds by rater

These results show that the substantial rater training and standard setting efforts did not suffice to achieve high (or even satisfactory) rater agreement, homogenous rating criteria interpretation and uniform PRLC-R use. Stark variability is commonly reported in research on rater-mediated performance assessments (cf. Eckes, 2005; McNamara, 1996). The present assessment is of low to no consequence to the test takers and thus, carries no implications for the stakeholders involved. However, the results are problematic nevertheless, and should similar instruments be implemented in more high-stakes contexts, it is indispensable to consider and mitigate the implications on test takers resulting from such stark rater variability (Eckes, 2005). Should this be the case, substantial adaptations to the rating scale would need to be performed and more research would be necessary before it could be used as a basis for making high-stakes decisions.

Rater Fit

After assessing the global model fit, rater fit statistics reveal more detailed information on individual rater behaviour. Here, rater fit indicates the degree to which a rater is associated with unexpected rating behaviour over the criteria and tasks (Eckes, 2005, p. 209). As introduced above, *rater infit* and *rater outfit* constitute two mean-square statistics that indicate data–model fit for each rater. The values from the table below are interpreted according to Linacre's (2002) suggestion to use 0.5 as a lower and 1.5 as an upper limit for infit and outfit mean-square statistics. Table 12 below shows medium to high outfit values, indicating that the ratings contain outliers and that the raters' behaviour clearly deviates from the behaviour expected. Because the infit statistic is a weighted fit, the values deviate less from 1.0, however they still indicate unexpected rater behaviour:

parameter	*N*	*Outfit*	*Infit*
rater 1	1036	3.085	0.972
rater 2	1090	1.187	0.624
rater 3	1442	1.422	0.629
rater 4	721	3.480	1.229

Table 12 : Rater outfit and infit statistics

When looking at the individual raters, the outfit values show that rater 2 and rater 3 evaluated performances more conservatively, and rater 1 and rater 4 judged more randomly. When interpreting these values, it is important to keep in mind that they stand in relative relation to one another. Similarly, infit and outfit mean-square statistics summarised over test tasks were computed resulting in values as shown in Table 13:

parameter	*N*	*Outfit*	*Infit*
task 1	574	11.136	9.394
task 2	616	0.688	0.654
task 3	663	0.672	0.665
task 4	637	0.622	0.600
task 5	637	0.798	0.627
task 6	616	0.805	0.713
task 7	546	0.709	0.668

Table 13 : Rater outfit and infit statistics summarised over test takers and test tasks

Low outfit or infit values (< 1.0) indicate that raters show less variation in their ratings with reference to test task than expected by the model; thus, both the outfit and infit values here overfit the model with exception of task 1. In task 1, raters behaved with substantially more variation than expected by the model; the value thus significantly misfits (i.e. underfits) the model. Outfit values that fall outside the 0.5 to 1.5 fit range are, according to Linacre (2002), less problematic than overly large (or small) infit values; however, misfit is considered more precarious than overfit (Myford & Wolfe, 2003). Accordingly, task 1 with a substantial misfit stands out as a particularly problematic test item. Overall, the values indicate relatively conservative rating behaviour across test tasks, i.e. more pronounced use of the performance levels 2 and 3 in the middle of the scale. Finally, Table 14 below shows the rater infit and rater outfit statistics summarised over rating criteria:

parameter	*N*	*Outfit*	*Infit*
accuracy	613	1.846	0.727
addressee-specificity	612	2.239	0.929
coherence & cohesion	612	1.573	0.844

content	613	2.317	0.803
fluency	613	1.670	0.687
pronunciation	613	2.641	0.798
vocabulary	613	2.611	0.896

Table 14 : Rater infit and rater outfit statistics test takers and rating criteria

These values indicate the extent to which the application of the assessment criteria deviates from what is expected by the model. The results show a pronounced *misfit* i.e. underfit (> 1.0) in the outfit statistics, indicating that they are problematic. As the rater outfit statistic is sensitive to individual unexpected ratings, the values above indicate that the raters show more variation in their ratings than expected by the model. The infit statistic reported above is sensitive to large amounts of unexpected ratings. With all values being below 1.0, they indicate an overfit of the model, thus indicating a halo effect among individual raters in the application of the PRLC-R criteria. In other words, the raters applied the rating criteria interdependently from one another despite rigorous rater training efforts.

Bias Analyses and Two-Way Interactions

This subsection reports on the bias and interaction analyses that were conducted to investigate rater behaviour and variability with reference to rater severity or leniency. Bias analyses were conducted (1) to examine individual rater severity or leniency across all tasks and all criteria, (2) to investigate overall rater severity or leniency per individual criterion, and (3) to test overall rater severity or leniency per individual task. Two-way interaction analyses were performed to examine if the individual raters judged with a consistent level of severity across individual assessment criterion (Rater x Criterion) and individual assessment task (Rater x Task). Thus, the analyses examined if the combination of a specific rater and a specific criterion, or a specific rater and a specific task, resulted in scores that were too severe or too lenient, respectively (Eckes, 2005).

Bias Analysis: Individual Rater Severity across all Tasks and Criteria

The table below outlines the modelled rater severity to answer question #2.1: Do the raters differ in the severity or leniency with which they rate the test takers' performances? The model shows that R1 and R3 were similarly lenient (0.137 and 0.103, respectively), and that R2 rated much more severely (0.366) across all tasks and criteria. R4 stands out as the most lenient with a value of -0.331:

parameter	*facet*	*xsi*[26]	*se.xsi*
rater 1	rater	-0.121	0.026
rater2	rater	0.333	0.023
rater3	rater	0.085	0.021
rater4	rater	-0.298	0.041

Table 15 : Rater severity examined through two-way interaction analyses

Bias Analysis: Overall Rater Severity per Individual Criterion

This analysis served to answer question #2.2: Do the raters maintain a uniform level of severity or leniency across criteria, i.e. do ratings on one criterion follow a pattern that is markedly different from ratings on the others (cf. Eckes, 2011, see chapter 5.1.2)? Table 16 below indicates the relative difficulty for an examinee to score highly in a given criterion. The closer the value is to 1.0, the more difficult it is for a test taker to receive a high rating for the respective criterion. The results indicate that receiving a high score in *content*, *pronunciation* and *fluency* was relatively easy. To score highly in *accuracy* and *addressee-specificity* falls somewhere in a medium-range of difficulty, and to receive a high score in *vocabulary* and *coherence & cohesion* was the most difficult. In other words, overall, raters were less likely to award a high score, i.e. were stricter when judging the criteria *accuracy* and *addressee-specificity*, and were even less likely to award a high rating when evaluating *vocabulary* and *coherence & cohesion*.

parameter	*facet*	*xsi*	*se.xsi*
content	item	0.299	0.041
vocabulary	item	0.710	0.039
accuracy	item	0.460	0.039
pronunciation	item	0.378	0.040
fluency	item	0.355	0.040
coherence & cohesion	item	0.746	0.041
addressee-specificity	item	0.471	0.039

Table 16 : Criteria difficulty examined through two-way interaction analyses

Bias Analysis: Overall Rater Severity per Individual Task

This analysis aimed at answering whether the raters' level of severity or leniency remained consistent across the seven test tasks. The below Table 17 shows the relative difficulty of a task

[26] *xsi* equals the estimated parameter / interaction estimate (which in this case relates to difficulty; the closer the value is to 1, the more difficult it is for a test taker to receive a high rating in the respective task or item i.e. criterion), and *se.xsi* indicates the measurement error; the closer the value is to 0, the lower the measurement error and thus the more reliable the measurement overall.

as judged by the raters. The results indicate that task 1 stands out for being much less difficult in comparison to the other six task, which were scored similarly harshly by the raters:

parameter	*facet*	*xsi*	*se.xsi*
task1	task	-1.497	0.037
task2	task	0.244	0.027
task3	task	0.332	0.026
task4	task	0.196	0.027
task5	task	0.176	0.027
task6	task	0.234	0.027
task7	task	0.314	0.070

Table 17 : Task difficulty examined through two-way interaction analyses

Congruent with the mean-square fit statistics described above, test task 1 stands out as being different from the others and thus problematic. Aside from problematic rater behaviour in task 1, one of the reasons for the problematic values could be construct-irrelevant easiness of the task (Xi & Sawaki, 2017). Task 1 itself instructed test takers to encourage a fictional student to participate in the classroom and at the same time to justify the need to learn the present perfect aspect in English. While the encouragement-part of the task directly reflected the construct, the grammar-related explanation did not. It could be that the participants were overall very familiar with the grammar rules surrounding the present perfect and thus found it particularly easy to complete the task. However, this result is difficult to find a plausible explanation for. In sum, the raters' behaviour (leniency or severity) across tasks shows rater variability with reference to rater-task interaction – especially in relation to task 1. Indeed, the findings show that the raters did not rate consistently with reference to severity or leniency across individual task as an overall group.

Two-Way Interaction Analysis: Rater x Criterion

This two-way interaction analysis was conducted by means of marginal maximum likelihood to measure "whether the combination of a particular rater and a particular criterion, or task, resulted in too harsh or too lenient scores awarded to some examinees" (Eckes, 2005, p. 213). A significance test can be conducted by dividing the interaction estimate by its standard error. If a Z score >|1.96|, the null hypothesis (here: the combination of raters and criteria, or tasks, did not result in scores that deviated from a uniform level of severity or leniency) is rejected.

The following Figure 22 displays differential item-rater functioning across raters and the PRLC-R criteria:

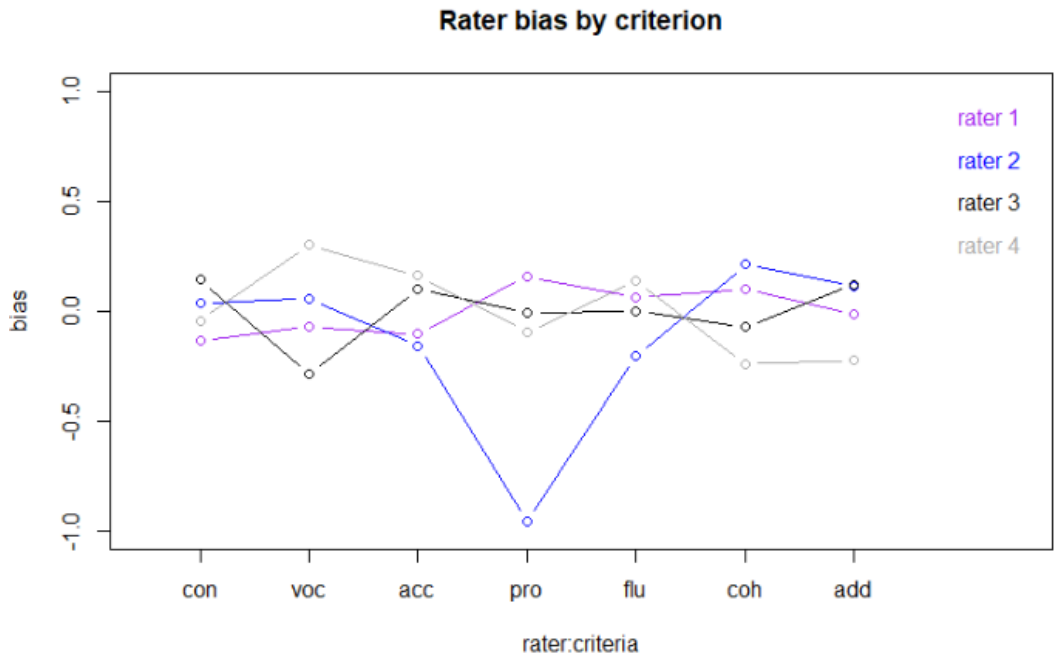


Figure 22 : Two-way interaction analysis Rater x Criterion

It is evident that raters did not employ the criteria in a uniform manner. Indeed, all raters display statistically significant variability in at least three criteria and R1 and R2 stand out with the most statistically significant variability across employing criteria and awarding scores: R1 *(content: Z = -2.52*, accuracy: Z = -2.08*, pronunciation: Z = 3.16*, coherence & cohesion: Z = 2**), R2 (*accuracy: Z = -3.721*, pronunciation: Z = -21.772*, fluency: Z = -4.591*, coherence & cohesion: Z = 4.711**), R3 (*content: Z = 3.55*, vocabulary: Z = -7.333*, accuracy: Z = 2.564**) and R4 (*vocabulary: Z = 4.054*, accuracy: Z = 2.145*, coherence & cohesion: Z = -3.064**).

Two-Way Interaction Analysis: Rater x Task

Another two-way interaction analysis was conducted to examine potential differential rater functioning between rater and task. Figure 23 shows stark variability in terms of rater-task interaction, indicating that there was almost no uniformity maintained within the ratings across raters and tasks (note the difference in scale in comparison to Figure 22):

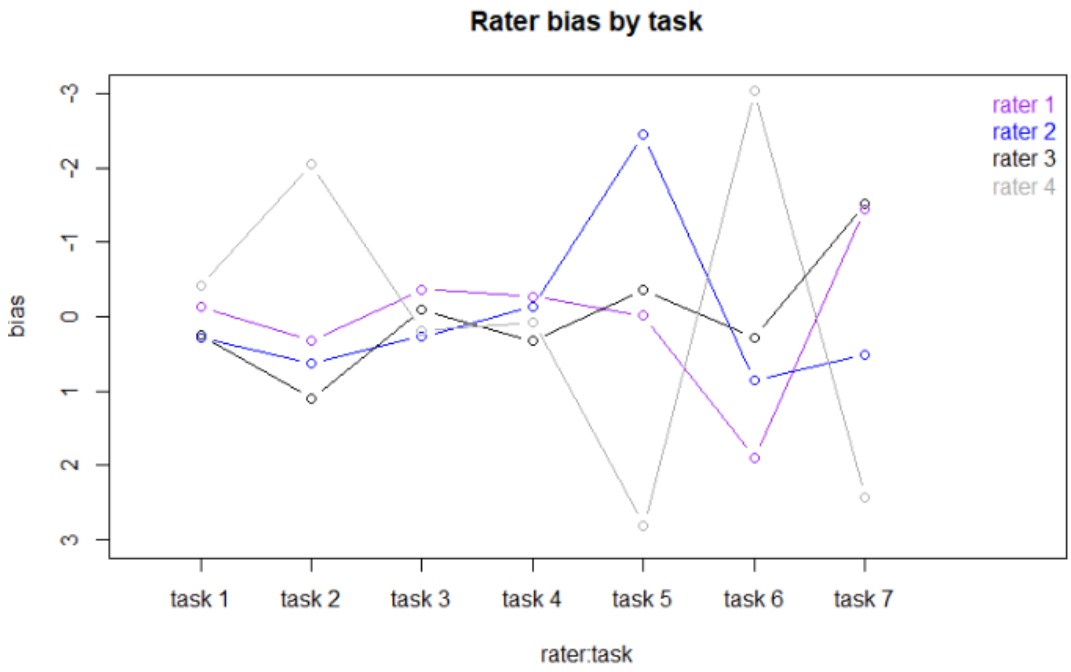


Figure 23 : Two-way interaction analysis Rater x Task

The calculated *Z*-scores indicate that the individual rater variability per individual task was in almost all cases (except for rater1:task1, rater1:task5, rater3:task3 and rater4:task4) statistically significant.

Gender Bias

To answer question #2.3 (see chapter 5.1.2) and uncover any potential gender bias in raters' evaluations, another *Z* statistic was modeled. The results show that rater 1 treated male and female responses similarly. Rater 2 and rater 4 judged male responses slightly more severely to a more or less equal extent, and rater 3 scored male responses significantly more severely:

parameter	*facet*	*xsi*	*se.xsi*	*Z score*
rater1:male0	rater:male	0.034	0.026	1.307
rater2:male0	rater:male	-0.052	0.023	-2.260*
rater3:male0	rater:male	0.119	0.021	5.666*
rater4:male0	rater:male	-0.101	0.041	-2.463*
rater1:male1	rater:male	-0.034	0.026	-1.307
rater2:male1	rater:male	0.052	0.023	2.260*
rater3:male1	rater:male	-0.119	0.021	-5.666*
rater4:male1	rater:male	0.101	0.041	2.463*

Table 18 : Gender bias *Z* statistic

Overall, raters (except for rater 1) awarded more high scores to female test takers (*male0*) in comparison to male test takers (*male1*), i.e. according to the statistics, female test takers scored

more highly than male test takers in the overall test. Finally, the model shows that females scored more highly by > 1 logit. This finding indicates some evidence that raters did not maintain a uniform level of severity or leniency across male and female test takers and that thus provides grounds for assuming that some gender bias is prevalent. Another explanation may be that female test takers did indeed perform better than male test takers.

Answers to Research Questions #2.1-2.3

Based on the above results, the research questions #2.1, #2.2 and #2.3 all reveal moderate to severe differential rater functioning including prevalent rater biases and rater effects. Specifically, the research questions can be answered as follows:

RQ #2.1 Do the raters differ in the severity or leniency with which they rate the test takers' performances?

a) Does each rater maintain a uniform level of severity, or do particular raters score more harshly or leniently than expected?

The raters showed overall stark variability in terms of severity and leniency in their ratings and they thus did not maintain a uniform level.

RQ #2.2 Do the raters maintain a uniform level of severity or leniency across criteria and across tasks?

b) Do ratings on one criterion follow a pattern that is markedly different from ratings on the others?

Stark variability in terms of severity and leniency could be observed per individual rater across all tasks and criteria, per individual rater and individual task, and per individual rater and individual criterion.

RQ #2.3 Do the raters show evidence of differential rater functioning related to test takers' gender; that is, do they maintain a uniform level of severity or leniency across male and female test takers?

The results indicate that R2, R3 and R4 show evidence of gender bias when scoring test performances, with the tendency to award higher scores to female test takers as opposed to male test takers.

5.2.3. Results Pre-Post Test

To compare the pre- and post-test results to identify possible treatment effects and to answer RQ #1, a competence comparison was conducted by means of a partial credits model (PCM). Competence differences were investigated within and between groups E, C1 and C0 according to t0 and t1. The differences were investigated with reference to test takers' overall performance across all PRLC-R criteria and possible competence development per individual rating criterion. The latter was conducted to identify the behaviour of qualitative, language-specific aspects of the test takers' oral feedback performances between t0 and t1. This section presents the results to the following research question and corresponding hypothesis:

Research Question #1:

> How do qualitative, language-specific aspects of pre-service English teachers' oral feedbacks in the target language English provided to lower secondary school students develop under the administration of a profession-related assessment rubric and systematic feedback training?

Hypothesis #1:

> Through the iterative and repeated application of the PRLC-R and peer feedback, the research participants' oral profession-related language competences improve as measured against the PRLC-R criteria.

A competence comparison of test takers' overall language performance between the pre- and post-test results within and between the experimental group (E), control group 1 (C1) and control group 0 (C0) was conducted. The group comparisons were calculated with estimated means and by means of WLEs (cf. Warm, 1989) with no adjustment for multiple testing. For these calculations, the data from the group C1 at t0 were used as the overall reference point to determine any treatment effects. First, overall and criteria-specific competence comparisons were performed to identify the competence differences of individual treatment groups at t0 and t1. The following table indicates the overall competence difference between groups at t0:

Groups	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
C1 vs. C0	0.403	0.3	42.7	1.345	0.1856
C1 vs. E	-0.0205	0.248	42.7	-0.083	0.9346
C0 vs. E	-0.423	0.307	42.7	-1.380	0.1747

Table 19 : Overall competence difference between groups at t0 (WLEs)

As can be seen from the table above, all groups performed overall at a similar competence level at t0. When investigating the overall competences between groups at t0 per individual criterion, the values paint a fairly similar picture. Indeed, the groups performed mostly uniformly across all individual criteria. It is only in terms of *accuracy* that the C1 group outperformed group C0 (p = .0327*) and group E (p = .0587*). Similarly, C1's performance with reference to *pronunciation* was better than C0's (p = .0277*) at t0:

Criterion	*Groups*	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
content	C1 vs. C0	1.22	0.841	66.2	1.488	0.1524
	C1 vs. E	-0.245	0.696	66.2	-0.351	0.7264
	C0 vs. E	-1.46	0.861	66.2	-1.697	0.0943
vocabulary	C1 vs. C0	-0.0605	0.443	70.5	-0.137	0.8916
	C1 vs. E	0.254	0.366	70.5	0.692	0.4911
	C0 vs. E	0.314	0.453	70.5	0.693	0.4906
accuracy	C1 vs. C0	0.74	0.339	67.4	2.181	0.0327*
	C1 vs. E	0.54	0.281	67.4	1.923	0.0587*
	C0 vs. E	-0.2	0.348	67.4	-0.575	0.5674
pronunciation	C1 vs. C0	1.1	0.489	64.6	2.252	0.0277*
	C1 vs. E	0.504	0.405	64.6	1.246	0.2174
	C0 vs. E	-0.597	0.501	64.6	-1.192	0.2378
fluency	C1 vs. C0	0.173	0.278	63.3	0.493	0.6235
	C1 vs. E	-0.0559	0.231	63.3	-0.243	0.8092
	C0 vs. E	-0.193	0.285	63.6	-0.678	0.5004
coherence & cohesion	C1 vs. C0	0.133	0.553	71	0.241	0.8193
	C1 vs. E	0.466	0.458	71	1.019	0.3116
	C0 vs. E	0.333	0.566	71	0.588	0.551
addressee-specificity	C1 vs. C0	-0.0882	0.443	64.6	-0.199	0.8428
	C1 vs. E	-0.521	0.367	64.6	-1.421	0.106
	C0 vs. E	-0.433	0.454	64.6	-0.954	0.3435

Table 20 : Competences at t0 *between* groups per criterion (WLEs)

The same overall and criteria-specific competence comparisons were calculated between groups at t1. Just like at t0, the all groups showed no significant competence difference at t1:

Groups	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
C1 vs. C0	0.523	0.3	42.7	1.746	0.0880
C1 vs. E	0.069	0.248	42.7	0.278	0.7821
C0 vs. E	-0.454	0.307	42.7	-1.480	0.1462

Table 21 : Overall competence difference between groups at t1 (WLEs)

When investigating the measured competences at t1 between treatment groups according to individual rating criterion, only the criterion *pronunciation* indicates a statistically significant difference between groups, namely C1 outperformed C0. This particular competence difference was already apparent at t0, thus rendering this finding a null-result:

Criterion	*Groups*	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
content	C1 vs. C0	0.148	0.841	66.2	0.176	0.8612
	C1 vs. E	0.267	0.696	66.2	0.384	0.7022
	C0 vs. E	0.12	0.861	66.2	0.139	0.8899
vocabulary	C1 vs. C0	0.556	0.443	70.5	1.256	0.2133
	C1 vs. E	0.254	0.366	70.5	0.694	0.4899
	C0 vs. E	-0.302	0.453	70.5	-0.665	0.5081
accuracy	C1 vs. C0	0.549	0.339	67.4	1.618	0.1102
	C1 vs. E	0.185	0.281	67.4	0.659	0.5519
	C0 vs. E	-0.364	0.348	67.4	-1.047	0.2987
pronunciation	C1 vs. C0	1.29	0.489	64.6	2.644	0.0103*
	C1 vs. E	0.466	0.405	64.6	1.151	0.2540
	C0 vs. E	-0.826	0.501	64.6	-1.651	0.1036
fluency	C1 vs. C0	-0.082	0.278	63.3	-0.294	0.6794
	C1 vs. E	0.104	0.231	63.3	0.451	0.6536
	C0 vs. E	0.186	0.285	63.6	0.652	0.5168
coherence & cohesion	C1 vs. C0	0.178	0.553	71	0.321	0.7490
	C1 vs. E	0.139	0.458	71	0.304	0.7617
	C0 vs. E	-0.0382	0.566	71	-0.067	0.9464
addressee-specificity	C1 vs. C0	0.08	0.443	64.6	0.181	0.8572
	C1 vs. E	0.157	0.367	64.6	0.429	0.6696
	C0 vs. E	0.0772	0.454	64.6	0.170	0.8655

Table 22 : Competences at t1 *between* groups per criterion (WLEs)

Interesting to note here is that, with reference to *accuracy,* group C1 outperformed group C0 and group E at t0 (p = .0327* and p = .0587*, respectively, see Table 20), but not at t1 (see Table 22). Thus, while the control group C1 performed slightly better in *accuracy* at t0, the difference between E and C0 and C1 is not anymore statistically significant at t1. This finding could mean that either the experimental group E and control group C0 improved, or the control group C1 worsened over the treatment period. These insights need to be treated with caution and in light of the rater analyses presented above. The findings of C0 showing slight tendencies of performing marginally lower than C1 and E are plausible given the fact that the C0 group are MSc students. Throughout their teacher education degree at the PHSG, the MSc students

receive less L2 training in comparison to C1 and E, because the latter two are MA students with a language focus. While these findings indicate slight competence differences *between* the treatment groups at t0 and t1, they do not indicate any treatment effects, i.e. any competence development from t0 to t1 *within* the groups, as shown below. To identify any competence development, Table 23 below needs to be consulted which shows the results of the overall pre-post competence comparison between the comparison groups:

Treatment group	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
C1	-0.0792	0.0787	38.3	-1.007	0.3204
C0	0.0409	0.115	38.3	0.355	0.7247
E	0.010	0.0845	38.3	0.122	0.9037

Table 23 : Overall pre-post competence comparison between groups (WLEs)

No statistically significant competence development can be identified. One can thus assume that the competences of all treatment groups remained the same from t0 to t1. When computing the pre-post competence comparison per individual criterion across all groups, the results paint the same picture. Indeed, no statistically significant difference becomes apparent apart from the criterion *content* (see Table 24). It thus seems that there was either a memory effect since the participants completed the same test tasks in both the pre- and post-test, or that the treatment period caused some sort of learning effect when it comes to providing feedback or to solving such tasks and adhering to the (identical) task instructions.

Criterion	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
content	-1	0.473	38.3	-2.120	0.0406*
vocabulary	-0.12	0.269	38.3	-0.444	0.6596
accuracy	-0.0545	0.195	38.3	-0.279	0.7816
pronunciation	0.156	0.267	38.3	0.582	0.5638
fluency	-0.0671	0.149	38.3	-0.452	0.6539
coherence & cohesion	-0.425	0.339	38.3	-1.252	0.2183
addressee-specificity	0.169	0.242	38.3	0.698	0.4891

Table 24 : Pre-post competence comparison *across* groups per criterion (WLEs)

When investigating the results further, the pre-post competence measurement comparisons between groups per individual criterion show that it is only the C0 group that improved with reference to *content*. This is slightly unexpected because the C0 group did not participate in the treatment but solely completed the pre- and post-test. Otherwise, no statistically significant differences are observable, as shown in Table 25:

Criterion	*Group*	*estimate*	*SE*	*df*	*t.ratio*	*p.value*
content	C1	-0.725	0.522	38.3	-1.390	0.1725
	C0	-1.79	0.763	38.3	-2.350	0.0240*
	E	-0.213	0.56	38.3	-0.381	0.7056
vocabulary	C1	-0.428	0.297	38.3	-1.443	0.1572
	C0	0.188	0.434	38.3	0.433	0.6672
	E	-0.427	0.319	38.3	-1.341	0.1878
accuracy	C1	0.218	0.215	38.3	1.016	0.3160
	C0	0.0276	0.315	38.3	0.088	0.9304
	E	-0.137	0.231	38.3	-0.592	0.5576
pronunciation	C1	0.079	0.294	38.3	0.268	0.7899
	C0	0.27	0.431	38.3	0.628	0.5338
	E	0.0406	0.316	38.3	0.129	0.8984
fluency	C1	-0.0374	0.164	38.3	-0.228	0.8206
	C0	-0.257	0.24	38.3	-1.072	0.2905
	E	0.122	0.176	83.3	0.697	0.4901
coherence & cohesion	C1	-0.284	0.374	38.3	-0.758	0.4530
	C0	-0.239	0.547	38.3	-0.437	0.6646
	E	-0.611	0.402	38.3	-1.520	0.1367
addressee-specificity	C1	-0.254	0.367	38.3	-0.953	0.3465
	C0	-0.0859	0.39	38.3	-0.220	0.8269
	E	0.424	0.286	38.3	1.481	0.1469

Table 25 : Pre-post competence comparison between groups per criterion (WLEs)

No *p* value adjustment for multiple comparison testing was performed. The few significant *p* values thus need to be interpreted accordingly.

In summary, the treatment of the present dissertation did not result in any statistically significant competence development within any of the participants, both between and within groups, and with reference to individual PRLC-R criteria. Thus, the null hypothesis of Hypothesis #1 cannot be rejected. Research question #1 can thus be answered as follows: no significant effects can be observed on pre-service English teachers' oral teacher language competence by the example of oral feedback provision in the target language English under the administration of a profession-related assessment rubric and systematic feedback training.

6 Discussion Main-Study

After the presentation of the overall results in the previous chapters, I will now discuss the empirical findings with reference to the action-oriented approach to conceptualising teacher language proficiency and the development and assessment of profession-related language competences in the L2 teacher education context. The main purpose of this research was to investigate treatment effects of the PRLCP- and PRLC-R-based intervention study on pre-service teachers' scored speaking performances of providing feedback to lower-secondary school L2 learners. In alignment with the overall research questions, further examinations into performance scoring and rater variability were conducted to examine the usability of the obtained data. By employing an MFRA, rater and interaction effects could be investigated in more detail, which allowed for more profound insights on the functioning and application of the PRLCP and PRLC-R. These understandings contribute to a more informed discussion of the teacher language competence construct in applied contexts. The following sections provide elaborations on specific, predicating aspects of the overall main-study to locate and interpret the findings within the greater context of L2 teacher education and L2 teaching and learning.

6.1. PRLC-R and Rating

The analyses with reference to the functioning and application of the PRLC-R throughout the rating process were central and decisive to the overall research study. The findings of the IRR analyses and the MFRA (see chapter 5.2) reveal a) that all raters differed strongly in severity with which they judged participants' L2 speech performances; b) that all raters applied the PRLC-R inconsistently in relation to criteria, tasks, and test takers, and c) that three out of four raters showed a tendency towards gender bias. First, the observed mean α level for all four raters at α = 0.338 (poor agreement), the α values across all rating criteria for rater pairs (R2 and R4 α = 0.430 with the highest, and R1 and R2 α = 0.257 with the lowest level of agreement), and the α values across all rating criteria for individual raters (*fluency* α = 0.501 with the highest,

and a*ddressee-specificity* α = 0.0806 with the lowest level or agreement) highlight that raters did not judge reliably (see chapter 5.2.1). Technically, the data usability of the data for measuring and pre-post-treatment effects was compromised if not threatened, especially without the implementation of an MFRA. It is important to treat the interpretation of computed reliability coefficients with caution, and there is substantial research evidence that human raters are subject to many systematic sources of variability (see chapter 2.5.4.4). For example, rater background can be of significant influence and a possible source of bias when evaluating oral performance (Winke et al., 2013). Such biases can manifest themselves on a variety of levels. Duijm et al. (2018) for example found in a comparison of trained and non-trained rater judgements that raters with advanced linguistic knowledge paid more attention to *accuracy* while "laypeople" focused more on *fluency* (Duijm et al., 2018). The present study shows similar results. R1 and R3 with the most linguistic expertise and L2 teaching and learning experience judged *accuracy* and *fluency* more uniformly, while R2 and R4 with less experience displayed statistically significant variation when judging these criteria. In addition, the mode in which the raters receive the language productions can influence *how* raters evaluate, and how reliable their ratings are. For instance, studies on listening perceptions of L2 oral performance suggest that visual cues present an important source of information for listeners to rely on when understanding a spoken text (Burgoon et al., 2016; Raffler-Engel, 1980). Nakatsuhara et al. (2020) conducted a study comparing the ratings of audio-recorded speech productions with ratings of live and video-recorded performances. They found that the limitations of audio-only rating conditions might depress the assessment scores of test-taker performance. In addition, examiners who rated in the audio mode consistently gave harsher scores than those rating in the video and live mode. Finally, they found that raters noticed a comparable amount of negative linguistic features irrespective of whether they rated the video- or audio-recorded performances. One of the reasons for this may be raters' more restricted role in the non-live modes. Because in Nakatsuhara's (2020) study raters did not simultaneously fulfil the role of the interlocutor, they could fully devote their attention to details of the speech production, including negative features and features that might remain unnoticed in a live testing situation. The additional verbal report data Nakatsuhara et al. (2020) gathered indicated that

> visual information helped examiners a) to understand what the test-takers were saying, b) to comprehend better what test-takers were communicating using non-verbal means (e.g., smiling, (un)willingness), and c) to understand with greater confidence the source of test-takers' hesitation, pauses, and awkwardness. (p. 19)

Based on these findings the authors suggest that video-based ratings are more reliable for evaluating and reflecting test-takers' interactional competence than audio-based ratings. Similarly, and in the context of the present study, one may hypothesise that the lack of visual information influenced the raters insofar that they may have "over-focused" on particular aspects of the speech performances, i.e. that raters were thus subject to a pronounced halo effect. Such an over-focus may have resulted in wrongfully punishing aspects that may otherwise be considered natural to spoken language and conducive to ensuring successful communication (cf. Luoma, 2009). These considerations would also offer a plausible explanation for the differential rater severity across assessment criteria (see chapter 5.2.2):

parameter	*facet*	*xsi*	*se.xsi*
content	item	0.176	0.039
vocabulary	item	0.542	0.038
accuracy	item	0.324	0.039
pronunciation	item	0.246	0.038
fluency	item	0.209	0.039
coherence & cohesion	item	0.637	0.039
addressee-specificity	item	0.384	0.038

Table 26 : Differential rater severity across rating criteria: interpretation

As the above table shows, *vocabulary* and *coherence & cohesion*, and to some extent also *accuracy*, were judged most harshly by all raters. When consulting the PLDs of these criteria, it is not surprising that such an audio-mode-induced halo effect or "over-focus" may have resulted in harsher scoring. For instance, the PLDs for *coherence & cohesion* state:

Kohäsion & Kohärenz Sich sprachlich und inhaltlich zusammenhängend und strukturiert ausdrücken	Sie/er drückt sich **durchgehend nicht zusammenhängend** und **nicht klar strukturiert** aus. Allfällige sprachliche Mittel zur Verknüpfung der Äußerungen sind **unpassend**.	Sie/er drückt sich **gelegentlich nicht zusammenhängend** und **nicht klar strukturiert** aus. Sie/er verknüpft ihre/seine Äußerungen nur mit einigen wenigen sprachlichen Mitteln, die **teilweise unpassend** sind.	Sie/er drückt sich **grundsätzlich zusammenhängend** und **strukturiert** aus. Sie/er verknüpft ihre/seine Äußerungen mit einer **begrenzten** Anzahl von **passenden** sprachlichen Mitteln.		Sie/er drückt sich **durchgehend zusammenhängend** und **klar strukturiert** aus. Sie/er verknüpft ihre/seine Äußerungen **flexibel** und sicher mit **präzisen und passenden** sprachlichen Mitteln.
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**

Table 27 : PLDs for *cohesion & coherence,* see also appendix C

The audio-based rating mode results in the PLDs explicitly *instructing* raters to focus on markers that may be more characteristic of written than spoken language, and focusing raters' attention more on quantifiable aspects that essentially contradict the underlying construct and

purpose of L2 performance testing (Luoma, 2009; McNamara, 1996). In this particular category for instance, one such aspect that raters were instructed to pay attention to was the *amount* or *frequency* of cohesive devices i.e. linguistic competences used in a test response. The higher the amount, the higher a rater was to score the oral L2 production. While such quantifiable aspects may aid to reach higher rater agreement, they also do not necessarily represent the underlying construct. First, the frequent use of cohesive devices is more strongly associated with (more formal) written performances than speech productions (cf. Luoma, 2009). Second, an L2 learner may be less likely to understand a teacher's L2 production the higher the amount of cohesive devices (i.e., the closer the production to structures of written language, the more complex and less comprehensible the input, cf. Wulf, 2001). This also indicates that such criteria might be more validly judged with a global rather than analytical rating scale. Furthermore, most audio recordings did not last for more than 1:30 minutes. Given such little speech material to judge against a criterion that usually requires longer speech productions to make an accurate evaluation, it is likely that an "over-focus" on construct-peripheral aspects may have become even more pronounced. At the same time, raters could not access the test takers' non-verbal cues used to complement their message and ensure understanding – cues that are particularly important in settings like the L2 classroom. The tendency of over-focusing on construct-peripheral or even construct-irrelevant aspects becomes apparent in the rating manual and the criterion *coherence & cohesion*:

> Zudem gilt es sich des inhaltlichen Wiederspruchs dieser Komponente bewusst zu sein: Bei der Beurteilung muss einerseits auf Details wie den Einsatz von sprachlichen Mitteln (quantifizierbar) geachtet und andererseits das Gesamtbild bzw. den Gesamteindruck der Aufgabenlösung hinsichtlich Kohäsion und Kohärenz (nicht quantifizierbar) in Betracht gezogen werden («kann man der Aussage gut folgen?»). Diese Diskrepanz in einer Komponente zu vereinen ist eine Herausforderung – und trotzdem sind beide Aspekte hier ausschlaggebend. (Rating Manual, p. 15, see also appendix D)

According to the manual, thus, raters needed to make both an analytic *and* a holistic judgement simultaneously. This aspect alone indicates that the criterion *coherence & cohesion* is not clear-cut and that thus its application is challenging. It also indicates that the descriptor may be wrongfully conceptualised or incorrectly worded. Similar such PLDs and guidelines are true for the other PRLC-R criteria employed in the present study, such as for example *accuracy.* At first, *accuracy* might seem a more straightforward criterion to evaluate, however when looking at the rating manual, a possible over-focus as outlined above seems almost unavoidable:

> Diese Komponente dient der Beurteilung der sprachlichen Korrektheit entkoppelt von jeglichen anderen Komponenten. Bei der Beurteilung von Ungenauigkeiten und Fehlern wird nicht zwischen schwerwiegenderen und weniger schwerwiegenden Fehlern unterschieden. Zu beachten ist hier die relative Häufigkeit von Fehlern in Bezug auf die aktive Sprechzeit. Treten in einer Aufgabenlösung mit kurzer Sprechzeit gleich viele Fehler auf wie in einer Aufgabenlösung mit langer Sprechzeit, wird erstere tiefer beurteilt als letztere. Diese Einschätzung beruht jeweils auf dem Eindruck der jeweiligen Aufgabenlösung. (Rating Manual, p. 12-13, see also appendix D)

The explicit instruction to focus on the relative frequency of grammar errors is likely to have significantly increased the chance of halo effects (i.e. "over-focus") and provides a possible explanation for raters to have judged the criterion more harshly. At the same time, the manual instructs raters to simultaneously base their judgement on an overall *global* impression of the task response – an instruction that stands in contradiction with the requirement to focus on the relative frequency of grammar errors. Evaluating *accuracy* proved additionally challenging because its application was not always as straightforwardly distinct from other criteria as implied in the manual and the PLDs:

Sprachliche Korrektheit Sich sprachlich korrekt ausdrücken (Grammatik)	Sie/er macht **so häufig** grammatische Fehler, dass **durchgehend unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **häufig** grammatische Fehler, wobei **teilweise unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **manchmal** grammatische Fehler, wobei **grundsätzlich klar** ist, was sie/er ausdrücken möchte.		Sie/er macht nur **sehr selten** oder **gar nie** grammatische Fehler, die auffallen.
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**

Table 28 : PLDs for *accuracy*, see also appendix C

For example, questions arose with reference to a test takers' vocabulary use and how to evaluate *accuracy* if a language production was generally grammatically correct but semantically (or syntactically) inappropriate. In other words, a production could be of high accuracy but unclear due to semantical or lexical issues, in which case it would be unclear what performance level to assign the speech production to. Although the PLDs and the manual offer superficial "solutions" to such issues, a halo effect is likely to have occurred in most cases – as shown in the results of the interaction analyses that indicate that the criteria were not applied independently from one another (see chapter 5.2.2). Another problematic criterion proved to be *addressee-specificity.* Indeed, special emphasis must be put on this criterion to illustrate the challenges of the PRLC-R criteria for performance testing, especially with reference to

indigenous criteria. *Addressee-specificity* stands out with its Krippendorff's mean α value across all raters of 0.0853. This value indicates that there is a complete absence of agreement, i.e. there is no statistical relation in terms of how the raters applied this criterion. That this criterion is difficult to grasp becomes apparent when consulting the PLDs:

Adressatenbezug: Lernende Sich den Lernenden gegenüber verständlich ausdrücken	Ihr/ihm **gelingt es nicht**, die Sprache an die Lernenden anzupassen, um ihnen das Verständnis zu ermöglichen.	Ihr/ihm **gelingt es nur teilweise**, die Sprache an die Lernenden anzupassen, um ihnen das Verständnis zu ermöglichen.	Ihr/ihm **gelingt es grundsätzlich**, die Sprache an die Lernenden anzupassen, um das Verständnis zu ermöglichen.		Ihr/ihm **gelingt es gut**, die Sprache an die Lernenden anzupassen, um das Verständnis zu ermöglichen.
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**

Table 29 : PLDs for *addressee-specificity*, see also appendix C

Similar to *coherence & cohesion* as outlined above, the criterion encompasses a range of inherent discrepancies and overlaps. For example, it poses the challenge that *addressee-specificity* is indeed not clear-cut as it addresses a multitude of assessment criteria simultaneously. For example, to adapt one's language to the target group (i.e. comprehensible input, cf. Wulf, 2001) essentially involves a variety of strategies, such as making adaptations to the vocabulary used, slowing down or speeding up the articulation rate, focusing on one's pronunciation and potentially adapting certain pronunciation features, implementing supplementary non-verbal cues, etc. (cf. for instance Wipperfürth, 2005). The mediation strategies as outlined in the CEFR-CV (Council of Europe, 2018, 2020) provide potentially valuable considerations of possible aspects that may be involved with *addressee-specificity* (see chapters 8.2, 8.4 and 9.2). Evaluating a test response in a test setting such as the present and according to the criterion *addressee-specificity* also encompasses first- and secondhand-inferences that raters need to make. The secondhand inference entails that raters need to guess the level of L2 proficiency of a fictional group of L2 learners based on a brief description in the test task (e.g.: "3. Klasse Oberstufe, Sekundarschule, erweiterte Anforderungen"; see also appendix B). This in itself is challenging and a precise inference is virtually impossible – especially when considering the complexity of human interaction on the one hand, and the complex, highly dynamic and multifaceted nature of the (L2) classroom on the other (Caspari et al., 2016; Königs, 2010). With most certainty in a test setting such as the present, and based on raters' individual experiences, all of the raters may have a different idea of what such a group of learners and their individual language proficiency would constitute (e.g., what socio-economic background do the raters attribute to the class? Do they imagine a class from a school

located in an urban or rural, rich or poor, Swiss or foreign area? How much pluricultural and plurilingual variance do they consider the class to have?, etc.). The firsthand inference includes the need for raters to guess the extent to which a test taker succeeds in adapting her or his language to that fictional target group. At the same time and based on a short language sample, the rater needs to be able to interpret whether the performance shown constitutes an *intentional* adaptation of the test takers' language to the fictional addressee, or whether it displays a test takers' *actual* language ability with no adaptations made. Both types of inferences essentially call on raters' field experience, diagnostic competence and prior knowledge rather than on concrete, distinct or quantifiable markers of language and/or assessment criteria. This makes *addressee-specificity* a highly elusive criterion. This elusiveness is emphasised in the rating manual, which states that

> [k]ommen […] Merkmale [nicht-adressatengerechter Sprache] in Aufgabenlösungen vor, muss jeweils abgewogen werden, inwiefern sie für die Lernenden auf der Zielstufe zu anspruchsvoll sind und Verständnisprobleme verursachen können. (Rating Manual, p. 16, see also appendix D)

Even though the rating manual contains a list of *non-addressee-specific* aspects of language (e.g., high articulation rate, slang, jargon, idioms, phrasal verbs, complex sentence structures, etc.), the coherent, clear-cut and consistent application of this criterion is highly complex – if not in reality entirely impossible. This complexity also becomes apparent in the post-rater-training questionnaire that was administered as a form of summative evaluation. The raters indicated that they had perceived the boundaries between the criteria and PLDs as blurred and fuzzy. In particular, two of three raters indicated that the criterion *addressee-specificity* proved to be the most difficult to grasp and judge accurately. This fuzziness is reflected in the results of the present data analyses (see chapter 5.1.2). A similar finding was obtained in a factor analysis that Bleichenbacher et al. (2017) conducted when investigating the distinctiveness of items derived from the PRLCP that had been implemented in a self-assessment tool:

> Eine zusätzliche explorative Faktorenanalyse für den Bereich Sprechen zeigte ebenfalls auf, dass die Anpassung der Sprache an die Lernenden sowie die Aussprache hauptsächlich eigene, von den anderen sprachlichen Fertigkeiten unabhängige Faktoren bilden. Dies legt den Schluss nahe, dass in der Befragung nicht nur Selbstbeurteilungen von berufsspezifischen Sprachkompetenzen als ein einheitliches Konstrukt abgefragt wurden, sondern dass sich davon die Selbstbeurteilungen zur Aussprache oder zur Fähigkeit, die Sprache an das Zielpublikum anzupassen als separate Bereiche abtrennen

> lassen. Die Ergebnisse deuten darauf hin, dass es sich bei der Selbstbeurteilung von berufsspezifischen Sprachkompetenzen um ein Konstrukt mit mehreren, voneinander zu differenzierenden Facetten handelt. Die empirischen Befunde der Skalierung zeigen, dass sich insbesondere die Selbsteinschätzung der Aussprache sowie eine angemessene Adressatenorientierung abgrenzen lassen. (p. 29-309)

These findings stem from analyses conducted with self-assessment items in relation to the PRLCP, however the underlying constructs correspond to those employed in other-assessment contexts that use the PRLC-R such as the present test setting. Hence, these findings are transferrable to this research context. That *addressee-specificity* may constitute its own, separate construct is not only indicated in Bleichenbacher et al.'s (2017) empirical analyses, but also in the present evaluation process as well as the additional, qualitative insights gained through the sub-study (see chapter 8.2). Contexts shaped through the PRLCP and PRLC-R alone, in combination with the complex nature of performance testing, render judging L2 teacher language competence by means of the PRLC-R a cognitively highly demanding and practically very challenging endeavour. It is thus not surprising that such complex cognitive processes involved with evaluating performance criteria, and inherent discrepancies within the criteria themselves, result in low IRR values. On the contrary, low rater agreement is in line with findings from the literature that show that even the most comprehensive of rater trainings, repeated double ratings and the computation of reliability coefficients do not offer any guarantee for reliable and consistent rater behaviour – especially not when it comes to assessing fuzzy and elusive constructs (Eckes, 2005, 2011; McNamara, 1996). That raters did not function interchangeably and did not rate homogenously throughout the rating process thus aligns with related research findings on the degree of severity exercised in language performance assessments (Eckes, 2005; Engelhard, 2002; McNamara, 1996; Myford & Wolfe, 2003). At the same time, both the IRR results as well as the unsatisfactory overall model fit do show that the rating criteria of the PRLC-R need to be optimised by conducting more research, refining them by implementing the findings, making them more clear-cut and thus ease their applicability and use (see chapter 6.7).

6.2. Pre- and Post-Test

The pre- and post-test analyses based on the performance ratings showed that there were no observable significant treatment effects on pre-service English teachers' oral teacher language

competence. Aside from the reliability concerns and problems related to the expert performance ratings that rendered the data largely unusable for conducting any competence comparison analyses and drawing any sound conclusions, a range of other reasons may have contributed to the lack of effects. A variety of those concerns the pre- and post-test. First, the pre- and post-test constitutes a weak LSP performance test that was developed for the local purpose of the present intervention study. The assessment instrument is based on a test construct (PRLCP) and assessment criteria (PRLC-R) that aim to conceptualise and operationalise the complex and multifaceted construct of teacher language competence. On the one hand, these foundational instruments present a much more distinctive conceptualisation of the teacher language competence construct that is, through its needs- and action-oriented approach, much more closely aligned with the demands of the real world classroom (cf. Kuster et al., 2014, see also chapter 2.3.3). On the other hand, as shown in the analyses regarding the application and interpretation of the PRLC-R above (see also chapters 5.1, 5.2, 8 and 8.2), the construct per se seems to still be far from being well-defined (cf. Elder, 2001). The fact that both instruments were not yet empirically validated at t0 and t1 may be a possible explanation for the prevailing elusiveness and fuzziness of the construct, and thus, for the lack of discriminatory power of the instruments when applied in performance assessment. In addition, the complexity of the construct of teacher language competence renders assessing language performance in the TLU domain – an already very complex type of language assessment – very difficult. As mentioned above, the pre- and post-test is considered a weak LSP performance test (McNamara, 1996). LSP tests usually encompass overlapping or *double constructs* and specific relations between language knowledge and content knowledge in the TLU domain (Douglas, 2000). Such coexisting *double constructs*, as is generally the case with LSP or LAPP performance tests, pose serious validity concerns. In this particular case, the pre- and post-test assess both L2 language proficiency required for teaching (i.e. classroom communication) and the LSP construct of teacher language competence. Overlapping constructs such as these make the present assessment instrument vulnerable to construct underrepresentation and construct irrelevant variance (Hoekje, 2016; Messick, 1994). As is common for weak performance tests, the present test draws on *TLU domain* constructs ("language" constructs), but the PRLC-R criteria – aside from *addressee-specificity* – underrepresent relevant aspects of teacher language performance (Hoekje, 2016). One such example is the fact that test takers could not make use of – or at least not show – paralinguistic cues to communicate, which in the real-world classroom would constitute a central component of teacher talk, and that there were no

corresponding PRLC-R criteria that raters could have used. Furthermore, even though the LSP test contains genuine stimuli and (near)authentic test tasks which increase the tests overall level of authenticity, there is no guarantee that profession-specific knowledge enables test takers to perform better on a profession-related language test than other highly proficient language users, which thus again threatens the discriminatory power of the overall test (Laurier & Baker, 2015). It is thus very likely that the present test does not suffice to draw meaningful conclusions about performance in the target domain, and even if the validity and reliability concerns were reduced to a minimum, that the test scores would need to be complemented with additional information like for instance some form of direct testing (Laurier & Baker, 2015). Finally, and in relation to the performance ratings, the results suggest that the assessment criteria did not discriminate appropriately between different performances (see chapter 6.1 for a more detailed discussion on the PRLC-R criteria and the rating process). It is also possible that the test tasks were too easy for the participants, which is indicated through the observable ceiling effect in the expert ratings (see chapter 5.2.1). Two explanations may be hypothesised: either, all participants already were at a high level of teacher language competence at the outset of the study, or they all were a similar level of competence both at t0 and t1. While all these reasons seem plausible, it is likely that there is not just one aspect that is responsible for the outcome, but rather the accumulation and interaction of all of the above – likely also in combination with aspects connected to the effectiveness (or lack thereof) of the BA E-Portfolio format on pre-service teachers' L2 teacher language competence development. The next subsection discusses such considerations and outlines potential reasons for the lack of treatment effects throughout.

6.3. Effectiveness of the BA E-Portfolio Format

In contrast to a range of studies that have shown that the combination of multi-stage assessment formats such as e-portfolios with (peer) feedback or other forms of reflective practice can positively affect learners' (self-perceived or other-assessed) L2 speaking skills (Bower et al., 2011; Cabrera-Solano, 2020; Castañeda & Rodríguez--González, 2011; De Grez et al., 2009; Gómez Sará, 2016; Hung & Huang, 2015; Kennedy & Lees, 2016; Lao-Un & Khampusaen, 2018; Murillo-Zamoranoa & Montanero, 2018; Yeh et al., 2019), the present intervention did not yield any significant treatment effects. As is known from the literature, the effectiveness of portfolios on learning achievement depends on a range of factors (Gläser-Zikuda et al., 2020). For example, research has shown that learners generally show a low level of acceptance of portfolios because of its high workload, the lack of relevance to learners' own needs and

interests, and the lack of objective evaluation (Gläser-Zikuda et al., 2020). In addition, learners' competence and skills of how to use the instrument effectively may also strongly influence the effectiveness of portfolios in L2 education. Indeed, learners' ability to adequately interact with and use a portfolio are preconditions for releasing portfolios' full potential (Ntuli et al., 2009). One possible explanation for why the intervention did not yield any statistically significant results may thus be that the participants may have had a critical attitude towards the portfolio task or that they lacked the necessary skills to achieve learning results. As no qualitative data on participants' perceptions has been collected, one may however only hypothesise. Trained and iterative peer feedback practice constitutes a further main component of the present e-portfolio-based treatment. A possible reason for the lack of treatment effects may be that the treatment itself was not sufficient or intensive enough to yield any observable results. In addition, it may well be that the treatment context has contributed to the lack of effects. Thus, it is plausible that practising teacher language competence with peers may not be as effective as for instance getting real-world TLU domain exposure and practising these competences in the actual L2 classroom. One may thus consider conducting follow-up replication studies that implement similar interventions, which for instance place the treatment in the context of pre-service teachers' practical placements, conduct more intensive feedback and oral teacher language competence practice (e.g., through peer feedback provision) in the relevant real-world TLU context, devise more rigorous feedback training in lectures, or even combine all three. Another possible reason for why the participants did not make any significant improvements on language-specific aspects of their feedback-related L2 teacher language competences might be that they did not trust the peer feedback they received. It is a common and stable finding that learners of any level generally prefer their teachers' or lecturers' feedback over their peers' feedback, and that they are consequently more likely to act on the former rather than the latter (Nelson & Carson, 1998). The participants may thus not have exploited the received feedback's full potential. Again, due to the lack of introspective data, such reasons can only be assumed. Another possible cause may have been participants' potentially low level of student feedback literacy. Low student feedback literacy results in learners not being able to recognise valuable feedback by their peers (Leki 1990; Stanley 1992) and thus not to revise their work accordingly (Connor & Asenavage, 1994; Liu & Sadler, 2003). Because the participants' feedback literacy skills were not measured, there is no imminent evidence that supports this hypothesis. Finally, the participants may indeed have provided low-quality, unusable and vague peer feedback. Such "rubber-stamp" advice has the potential to largely counteract the benefits of peer feedback

(Leki, 1990; Lockhart & Ng, 1993; Mendonca & Johnson, 1994; Min, 2016; Tsui & Ng, 2000) and thus prevent any possible treatment effects in the present intervention study.

Even though the implemented treatment did not show any statistically significant results, it would not do the study justice to claim that the BA E-Portfolio task adapted for the purpose of this dissertation did not lead to any positive results or learning insights. The adaptations made to the BA E-Portfolio format contributed to better structuring the overall task, to more explicitly guiding the participating students and to providing a more stable basis for the project group supervisors to accompany the students throughout the process. While the peer feedback component did previously not contain any explicit structure or explicit guidelines, implementing the PRLC-R as a basis for students to provide feedback delivered a research-and development-informed approach to refining the task and to aligning it more closely to the TLU domain. Despite the lack of introspective data, preliminary insights from the pilot-study support its favourable reception by the students (see chapter 4.2).

6.4. Creating a Validity Argument

According to Bachman (2004), test validation begins at the start of the test development when the test purpose, use, interpretations and consequences are determined. There are different approaches to verify the validity of a language test (see chapter 2.5.1.1). One of those is to take an argument-based approach and establish a validity argument (L. Bachman & A. Palmer, 1996; Chapelle et al., 2008; Kane, 1992; Kane et al., 1999). The scope and nature of examining test validity and the methods to gather evidence to support a sound validity argument depend on the way in which validity is conceptualised (Xi & Sawaki, 2017). For instance, validity evidence can be of empirical or theoretical nature, and it can be collected before (*a priori*) or after (*a posteriori*) the test administration (Weir, 2005). In addition, the rigour and elaborateness necessary to provide evidence of an instrument's validity depend on its purpose, context, scope, implications, stakes and stakeholders (Douglas, 2010). In the following sections, I discuss the validity of the present pre- and post-test and the PRLC-R. In contrast to examining test (and PRLC-R) validity on a piecemeal basis and investigating different types of validity separately, I loosely align the present approach to validity with Chapelle et al.'s (2008) framework, seeking to holistically integrate validity evidence into a coherent argument (see chapter 2.5.1.1). In an attempt to construct a coherent validity argument, I present theoretical and empirical validity

evidence collected before, during and after the test and PRLC-R development, and before and after the implementation of the intervention study.

6.4.1. Test Validity

To provide the formal scaffolding for constructing a coherent validity argument, I refer to the aspects Douglas' (2010) outlines that determine the rigour and elaborateness required to provide evidence of a test's validity. First and formally, the present test has no implications for the test takers. It was implemented for scientific research purposes only and the participants' test scores did not influence their marks, degree or subsequent professional endeavours. Within the formal structures the pre- and post-test is embedded, it can therefore be considered a low-stakes test. However, since the test instrument's purpose is to measure potential effects of the intervention design on the research participants' profession-related language competences, the validity of the test is an important requirement for drawing scientifically sound and empirically valid conclusions. Hence, some form of test validation is necessary. Test validation is a process that

> should be constantly ongoing – no test is ever validated once and for all time since as new populations take the test or as it is used for new purposes, new evidence must be marshalled to show that the interpretations made of test performance are justified [. Second], there is no such thing as a valid *test*, only tests which have been shown to be valid for certain purposes. (Douglas, 2010, p. 35)

Based on Weir (2005), I established an *a priori* argument for the content validity of the test tasks with reference to AoA 3 of the PRLCP. First, I supplemented the theoretical argument for the content of each test task with what Douglas (2010) calls *secondary* data. This included consulting the research literature and feedback from field experts (in-service teachers, language teaching and learning experts, linguists and language teacher educators), as well as language testing experts (see chapter 4.3.2.1). The *a priori* validity evidence collected from the consultations with the chosen experts, and thereby the inclusion of multiple perspectives, supports the content and face validity of the test. With reference to face validity, the test instrument requires authentic language-teacher-related tasks that focus on task completion and functional aspects of language (Laurier & Baker, 2015). With this collection of *secondary* data, I could confirm the real-world relatedness of the test tasks and the tasks' focus on task accomplishment. The underlying model of the pre- and post-test constitutes the PRLCP. The

profiles were developed based on an extensive needs analysis taking an action-oriented approach (see chapter 2.3.3), and even though the AoA including the individual descriptors were validated by means of several rounds of expert consultations, no empirical validation of the profiles per se has been undertaken to date. However, needs analyses are a common form of establishing the content relevance and test task representativeness to the target domain (Xi & Sawaki, 2017). Since the evidence that supports the domain description inference is in often based on (expert) judgements, one could claim that this process of the PRLCP development included important validation steps that now contribute largely to the validity argument of the profiles – and hence, the present pre- and post-test. Nevertheless, similar to the OPI that builds on guidelines that are experientially rather than theoretically based, the present test lacks an empirically validated competence model as its foundation. The PRLCP contextualise the descriptors of the CEFR, hence adopt the CEFR's model of communicative competence somewhat, but they do not have a contextualised competence model of their own. In contrast, the PRLCP are built on the results of an extensive needs analysis (Kuster et al., 2014; Long, 2005), which strengthens the face validity of the profiles. Still, Bachman's (1988) criticism of the OPI also applies to the current context:

> the ACTFL oral interview is based on procedures that have been developed over years of practice and about which a great deal of experience has accumulated. However, this experience in no way constitutes direct evidence for the validity of the test. (p. 160)

There are indeed validity concerns with reference to construct underrepresentation of the test because of the reduction of the richness of test takers' performances to audio recordings only. Because test takers are deprived of using paralinguistic cues to communicate their message – which in the real-world classroom would constitute a central component of teacher talk – the construct of teacher language competence is underrepresented. In addition, the present test is classified as a weak performance test. According to Douglas (2000), LSP tests "are characterized by an interaction between language knowledge and content knowledge in the specific domain" (Laurier & Baker, 2015, p. 22). With reference to this definition, and in the teacher education context, Elder (2001) points out that

> [t]eacher language proficiency [is] far from being a well-defined domain relying on highly routinized language and a generally accepted phraseology such as is the case with, for example, the language of air-traffic controllers. (p. 152)

The complexity of the construct of teacher language competence must not be underestimated, as it makes measuring language competence in this domain very difficult. Similarly, Laurier and Baker (2015) draw attention to the fact that, while implementing genuine stimuli and constructing (near)authentic test tasks contribute to strengthening the face validity of a test – just like in the present test (see chapter 4.3.2.1) – such measures do not guarantee that profession-specific knowledge enables test takers to perform better on a profession-related language test than other highly proficient language users. In sum, the interpretations of the test scores of the present pre- and post-test must be carried out with care and ideally be complemented with additional information such as, for example, some form of direct testing. As Laurier and Baker (2015) rightfully conclude, a single test does not suffice to draw meaningful conclusions about real-world performance.

6.4.2. PRLC-R Validity

Because assessment rubrics constitute a central component of language performance testing, it is important to assert that the criteria reflect the underlying skills that a test is supposed to measure (Xi & Sawaki, 2017). The validity of the respective rubrics can be tested, for instance, through rater verbal protocols (Brown et al. 2005), MFRA to determine whether differences between score categories are clear-cut (McNamara 1996), or multidimensional scaling when developing scales for different tests and raters (Xi & Sawaki, 2017). In case of the PRLC-R, multiple consultations with field experts and language teaching and language testing experts, and several iterative feedback and revision processes were undertaken prior to the present study. Within this study, pre-piloting and piloting the rubric with the target group (see chapter 4.4.1), consultations with the rating committee with subsequent adaptations, and rater questionnaires contribute to the validation argument. However, no statistical validation procedures and analyses were conducted prior to the implementation of the PRLC-R in the present study. The results obtained through interrater reliability, rater bias and interaction analyses by means of MFRA contribute to further validating the PRLC-R. In sum, the following steps undertaken contribute to constructing a validity argument:

- overall iterative development processes,
- pre-pilot and pilot studies for both the pre- and post-test as well as the PRLC-R,
- iterative consultation with experts and the subsequent revision of both instruments,

- extensive rater training with ample time for discussion and attempts to establish consensus,
- rater questionnaire,
- rating manual,
- statistical analyses to investigate systematic rater bias, interrater reliability and interaction effects.

6.4.3. Validity Argument: Verdict

Despite the above efforts to collect evidence to support a validity argument for the present test and PRLC-R use, not all interpretive arguments, inferences, and pertinent assumptions have been adequately addressed. Instead, this section presents much more a selective argument. As Kane (1992) and Xi and Sawaki (2017) rightfully caution, creating an argument that is driven by the availability of resources "may very likely have weak assumptions or even more seriously, weak hidden assumptions that are not even articulated in the argument" (Xi & Sawaki, 2017, p. 204). Thus, even though there is some compelling evidence that contributes to a satisfactory validity argument of both the pre- and post-test and the PRLC-R, weaknesses and loopholes cannot be excluded from the final validity declaration. Considering the stakes, scope and purpose of the present test and the PRLC-R, and considering that the generalised framework used (Chapelle et al., 2008) needed to be adapted to specific local context of present instrument use, however, one may argue that the argument presented in this section is fair and reasonable.

6.5. Limitations Main-Study

The above validity argument leads to the discussion of the overall limitations of the main-study. Generally, in (quasi-) experimental designs such as the present study, the independent variable (IV) is manipulated to investigate whether the treatment has any effect on the dependent variable (DV). In order to draw feasible inferences from the treatment, the following three requirements need to be met (cf. Darsow & Felbrich, 2014, p. 230):

1) There needs to be covariation between the DV and the IV; in other words, both variables must change in dependence of each other.
2) The change of the IV needs to happen before the DV changes.

3) Alternative explanations, caused for instance through confounding variables, need to be as implausible as possible and thus controlled.

To meet requirement 1) and 2), pre- and post-tests before and after any treatment are common procedures (Reimann, 2020a). Requirement 3) depends on the degree of the established internal validity of the research study design, for example by dividing the research sample into a control (C) and experimental (E) group (Darsow & Felbrich, 2014; Reimann, 2020a). To reduce confounding variables to a minimum, the allocation of the research participants to the control and experimental group should be randomised (Reimann, 2020a). In L2 teaching and learning research, however, this form of randomisation can pose a challenge merely due to the nature of the field itself. Randomisation of the research participants in the present study was only possible on the project-group level rather than the individual participant's level (see chapters 4.4 and 4.4.2). In addition, the participants constituted an entire cohort. Thus, the research sample can strictly speaking be considered an ad-hoc sample or a convenience sample because its selection was not based on theoretical sampling or sample theory. This sampling can have an influence on the quality of the results (Reimann, 2020a). The quasi-experimental nature of the study and the relatively small sample size (n=33) render the research a "weak" study (Darsow & Felbrich, 2014, p. 239) and do not allow for any generalising conclusions to be drawn. However, as the given sample represents an entire student cohort, and since this study serves predominantly explorative and descriptive purposes, conducting this study with a convenience sample is nevertheless reasonable (Grum & Legutke, 2016). Another common challenge of field research poses the dichotomy between internal (direct reference to constructs) and external (transferability to authentic real-world contexts) validity. While field research generally allows for higher external validity, such research designs often encompass cutbacks with reference to internal validity (Darsow & Felbrich, 2014). As Darsow and Felbrich (2014) explain:

> Ergebnisse einer Studie [sind] lediglich dann belastbar, wenn eine hohe interne Validität vorliegt und eine angemessene Veränderung eindeutig auf das *Treatment* zurückgeführt werden kann. […] Als starkes Design […] gelten experimentelle Versuchspläne […] Quasi-experimentelle Studien werden dagegen zu den schwachen Designs gezählt. (p. 239)

High internal validity, for instance, requires the use of standardised, valid, reliable and objective measurement instruments or verification of action steps taken to eliminate (the main) methodological biases (Petticrew & Roberts, 2006). While the validity of the research instruments are discussed above (chapter 6.4), I would like to touch upon aspects that threaten

the ecological validity and face validity of the present research instruments. With reference to the pre- and post-test administration, face validity and ecological validity become particularly interesting, especially when considered in combination with authenticity. Despite the attempt to increase the test's authenticity by implementing real-world, near-authentic stimuli, the overall level of authenticity during the test administration is still compromised. For instance, the test takers complete a near-authentic task including classroom talk while being seated in front of a computer and recording their responses by means of clicking a button and speaking into a microphone. This setting alone does not correspond to an everyday classroom situation, is however determined by the test administration and practicality reasons. In addition, the test environment of the pre-test differed from the post-test's. Due to the restrictions caused by the COVID-19 pandemic, the post-test fell into the government-issued lockdown in March 2020. In both the pre- and post-test, the participants completed the test on their own devices (BYOD). Instead of completing the post-test in a supervised condition like during the pre-test, the participants undertook the post-test from home during a pre-defined time slot. Despite all test takers signing a declaration in advance to certify that they would not use any unauthorised help during the test, there was no way in which this could be controlled. With this change of test-environment due to the recent COVID-19 pandemic, and a consequential global surge in distance learning and distance teaching, the compromised level of authenticity mentioned above no longer seems as severe during the post-test. In the distance-teaching format, it is not uncommon for teachers to sit in front of a computer and speak into a microphone when teaching. The physical environment is thus not so far removed anymore from an authentic distance-teaching situation. This aspect may become increasingly more relevant, depending on the long-term impact the COVID-19 pandemic on teaching and learning practices. Other limitations in connection with the pre- and post-test administration encompass the test setting in general. Because there were between two and five participants in each room when they completed the pre-test, it is likely that they might have influenced one another during the test completion. This influence may be twofold: either through being distracted by hearing other participants speak, or by hearing other participants solve a task and then copying the answer.

A further set of limitations concern the rating of the pre- and post-test data. The collected data consisted of audio-recorded test responses. First, raters (and research participants of the sub-study, see chapter 8.4 for limitations) could listen to the audio recordings more than once, which does not correspond to an authentic student-teacher-interaction. The perception of the situation was thus different from the classroom context, which means that the obtained results cannot be

transferred completely to the classroom. Indeed, excluding visual information from the rating process, or, in other words, reducing the available information to audio information only excludes nonverbal communication from the judgement process. However, nonverbal communication is highly important in professions that rely largely on interpersonal co-construction and negotiation of meaning such as teachers or medical professionals. Hoekje (2016) argues, for example, that the "use of eye contact and other nonverbal behaviors [is] essential to classroom communication" (p. 294). In research related to identifying indigenous criteria for the OET, for example, nonverbal behavior in clinical encounters was confirmed to be highly important (Elder & McNamara, 2016). However, the OET with its positioning as a language test that is subject to a range of constraints relies on audio recordings only for its ratings, just like the present test. This is a significant limitation because the underlying theory to language performance assessment conceptualises nonverbal communication as part of strategic competence (cf. Bachman & Palmer, 1996; Council of Europe, 2001). By removing this competence constituent, only partial judgements or inferences on a learner's communicative teacher language competence are possible. Likewise, the thus reduced level of authenticity compromises the test's ecological validity.

With reference to the quality and reliability of the ratings themselves, it is important to stress that rater biases could only be controlled to a limited degree (e.g., through the extensive rater training or the rating manual). Because rater bias – and rater effects, for that matter – are systematic and deeply rooted individual beliefs and behaviours, it is impossible to rid a rating process of them entirely (see chapter 2.5.4.4). By employing appropriate statistical analyses like an MFRA, rater effects could be compensated to a certain extent. Nevertheless, an MFRA does not eliminate a potential lack of rater consensus. Even though the extensive discussions during and after the rater training indicated that the raters had reached consensus – at least on a surface level – raters' underlying thought processes and latent convictions could not be accessed. In-depth qualitative retrospective interviews could reveal such insights; however, I only conducted a small-scale, informal and surface-level rater survey instead of interviews. Additionally, the mid-rating-process rater conference for additional consensus-seeking may have influenced the way raters judged the performances post-conference, especially because the ratings conducted prior to the conference were not readjusted after the consensus had been reached anew. This may mean that ratings that were allocated before the rater conference may differ from those allocated after the conference. While this may have contributed to potentially increasing interrater reliability, it may also have increased within-rater inconsistency – an

aspect that threatens the quality of the data and their usability for MFRA much more severely (see chapters 2.5.4.4 and 5.1.1). Finally, the methodological approach employed in this study grounds in social constructivist epistemology that views knowledge as being socially co-constructed and assumes the existence of numerous interpretations of reality (Crotty, 1998). Based on this approach, the "findings" presented in this research can only be seen as one interpretation out of many possible ones rather than the only interpretation or "truth". The results need to be interpreted accordingly. Taking into account these considerations, the present research design can in its essence not yield generalisable and conclusive scientific findings. Rather, it constitutes an exploration of the opportunities, challenges and affordances of the PRLCP and PRLC-R when applied in teacher education whose findings can be transferred to other comparable contexts.

6.6. Ethical Considerations Main-Study

I conclude this chapter with ethical considerations in relation to the overall main-study. The first consideration concerns informed consent of the research participants – a central concept in any quantitative or qualitative research study that includes human beings as research participants (Halse & Honey, 2005). Informed consent encompasses that the researchers provide full and accurate information about the respective study to autonomous subjects who, based on this information, are able to make rational, informed decisions on whether they want to participate in the research. All research participants of the main-study received detailed information on the present research, its procedure and its aims in informational e-mails and during the introductory session of *Introduction to Linguistics*. Upon each participation in the pre- and post-test, the research participants provided written consent directly via Moodle for their speech productions to be used for data analysis. Likewise, close attention was paid to concealing the participants' identity in the analysis, interpretation and discussion stages. All data was anonymised and randomised immediately after the respective data collections in preparation for the expert ratings. The second consideration concerns potential benefits or drawbacks that research participants could experience due to their participation in the study. Because group E and group C1 both received some form of further training in providing feedback and their treatment only differed in terms of the evaluation criteria they were asked to use, one can assume that neither of the conditions experienced any detrimental effects. It can thus confidently be claimed that all participants involved were treated equally and fairly and received some form of additional support – albeit slightly different.

6.7. Implications and Conclusions Main-Study

This dissertation study empirically investigated the use of the PRLCP and the PRLC-R in the Swiss L2 teacher education context. Overall, the implications with reference to the PRLC-R and the PRLC-P constitute that there is a strong need for them to be optimised. Such steps should include conducting more research into their application and functioning in L2 teacher education and language testing settings (see chapter 10.1), and refining the PRLC-R criteria according to the present research findings to ensure that they become more clear-cut and thus better applicable, more user-friendly, reliable, and of higher validity and systemic relevance. Based on the above discussion of the research findings, this section outlines implications in relation to the research instruments that seek to define, foster and assess teacher language competence (6.7.1), as well as to didactic consequences of the main-study findings both on the level of L2 teacher education and general L2 education (6.7.2).

6.7.1. Consequences for the Research Instruments

The main-study findings lead to a range of consequences that concern the PRLCP, PRLC-R and the pre- and post-test that have been implemented as research instruments in the present study. As with any development process, test- and scale-development follow an iterative and cyclical pattern (L. Bachman & A. Palmer, 1996; Harsch, 2016), for which Bachman and Palmer (1996) identify the following steps: (1) initial specification, (2) tryout, (3) analysis, and (4) revision (p. 362). In the present study, I conducted step two and three with reference to the PRLCP and PRLC-R, and step one, two and three with reference to the pre- and post-test. The results obtained from this study in relation to the PRLC-R can now be used for step four to revise the assessment rubric and the criteria, or overall to run further cycles of all four steps. Indeed, the findings on rater behaviour and rater effects should be used to inform the revision of the PRLC-R and related rater-mediated assessment in general. In other words, these insights should be treated and implemented as an integral source of information for further improving the psychometric quality of the present and similar tests using the PRLCP and the PRLC-R.

One measure to do so specifically concerns the individual PRLC-R criteria themselves. A striking problem of the PRLC-R is that it seems that the criteria are not clear-cut and that they do not assess the same underlying dimension, especially as long as the criterion *addressee-specificity* remains unaltered (see chapter 6.1). Indeed, this criterion seems to represent its own

distinct construct that diverges from the other formal linguistic criteria as they were rated in this study. This aspect alone calls for an overall revision of the PRLC-R assessment criteria, either through removing indigenous criteria entirely (and thus creating a more reliable *weak* performance test) or through further differentiating them (and thus seeking to develop a *strong* performance test with higher validity). The findings that indicate that some criteria cover construct-peripheral aspects point towards the need for revisiting the overall nature of the PRLC-R and for considering deriving a (potentially additional) global rating scale to more validly capture and assess teacher language competence. In addition, the consideration of adding additional criteria that cover diagnostic competence with reference to L2 learners' L2 proficiency, and paralinguistic features of teacher language competence (e.g., facial expressions, gestures and other strategies used to facilitate understanding) seems worthwhile – be this in form of an analytic or holistic rating scale component. This would enable assessors to make more comprehensive and valid judgements on L2 teachers' profession-related language competences and provide learners with a more holistic evaluation of their skills with a closer connection to the real-world TLU domain tasks – provided that learners can show and assessors can observe these features in the respective performances (i.e. in direct assessments or through video-task-responses and video-ratings). While such measures may contribute to higher ecological validity when it comes to measuring and assessing teacher language competence, they may at the same time further complicate the differentiation of the construct itself and the quest to ensure reliable rating procedures. Indeed, the multifaceted, complex and highly adaptive and situational nature of the teacher language competence construct, enriched by even more differentiated criteria as suggested, may render ensuring interrater reliability an almost impossible quest. This leads to the next implication of the research findings for the research instruments: consequences for applying the tools in related rater trainings.

The stark variability in rater functioning in relation to the assessment criteria and test takers indicate that within-rater consistency needs to be emphasises more strongly as opposed to between-rater consistency (Eckes, 2005), because ratings from internally consistent raters can more easily be used to conduct MFRA and compensate for rater effects. In order to be able to achieve higher within-rater consistency, the assessment criteria and the rater training procedures need to be further refined through taking evidence-based approaches. Aside from differential rater functioning in relation to assessment criteria, the inconsistency of the ratings across tasks highlights another precarious feature of the PRLC-R and the pre- and post-test. The overall rating process should therefore be revised in order to not only increase rater consistency

in relation to criteria, but also – and importantly so – in relation to tasks *and* test takers (Eckes, 2005). A global rating scale could contribute to achieving this goal. Furthermore, the pronounced differential rater functioning in relation to test takers' gender outlines the necessity to raise raters' awareness thereof through appropriate training. A possible gender bias is unlikely to be counteracted through better-refined assessment criteria, however appropriate training may be of benefit.

With reference to the PRLCP more specifically, the criterion *addressee-specificity* in particular needs to be further refined and better understood. One way of doing so may constitute conducting explicit adaptive-language and comprehensibility assessment of feedback conversations with pupils at the target level to identify whether the extent to which the teacher adapted their expression to the proficiency level of the learner is sufficient, situational and sensitive to context and content. By doing so, adaptive behaviour could become observable and thus would make it possible to be translated into descriptors and criteria and finally to be evaluated. Such procedures could then contribute to further differentiating the construct of teacher language competence, or specifically, the PRLCP (and PRLC-R) in general.

Finally, as implied above, it is worth considering the option of trialling video-recorded test responses as opposed to audio-recorded responses in tests that seek to measure teacher language competence. Video-recorded answers would allow for a more comprehensive evaluation of the competences in question and for less influence of construct-peripheral or irrelevant features when rating the respective performances. In addition, video-vignette-based competence-oriented performance tests could be further refined by including more innovative approaches to recording video-material. One such promising option constitutes the recording of video-vignettes by means of a 360-degree video camera and providing test takers with 3D goggles for completing the test tasks. Such an approach would then allow for a much higher degree of immersion in the TLU task on the test taker's side while maintaining comparability across test tasks (as opposed to direct testing in the classroom).

6.7.2. Didactic Consequences

A range of didactic consequences can be derived from the insights of the present main-study. One of them concerns the implementation of the research instruments in L2 teacher education. For instance, the developed pre- and post-test and adapted PRCL-R can be used as teaching materials or formative assessment tools. One such pilot was already conducted at the PHSG in

the autumn term 2020, where the pre- and post-test tasks were used as practice materials for training profession-related language competences in a French L2 module. In the last session of the term, the lecturer collected short feedback statements from the three students that were present. The participants provided the following responses (✓ equals positive feedback, *tableau* refers to the PRLC-R):

- ✓ le travail avec les vidéos est une bonne base pour les réponses qu'on doit enregistrer
- ✓ c'est un bon devoir, parce que c'est agréable de le faire à la maison et de plus ça va assez vite
- ✓ j'aime qu'on puisse vraiment utiliser ces compétences pour le travail
- ✓ Le programme […] est également très utile. Il permet de répondre spontanément aux réponses des élèves. Il est parfois difficile de donner un feedback aux autres étudiants parce qu'ils sont au même niveau.
- ✓ tableau […] nous donne une structure pour donner des feedbacks

It is especially the reported relevance of such tasks for real-world teaching practice that contributes to the tasks' positive reception among the participants. Implementing these tasks in teacher education allows for practising the relevant target language use in a safe environment with very low organisational and administrative effort. Such practice activities could be used in preparation for pre-service teachers' internships, or generally, to provide L2 teacher candidates with alternative options to rehearse classroom talk that is closely related to the authentic classroom. In addition to practising (more or less) general L2 teacher language use in a simulated TLU domain, implementing learning activities including these tasks can allow pre-service teachers to develop their feedback literacy. Indeed, providing pre-service teachers with explicit training in efficient feedback strategies, i.e. explicit training in order to develop both their student *and* teacher feedback literacy could be executed by means of implementing the present test tasks – ideally in multi-draft assignments that contain reflective components (cf. Yeh et al., 2019). Based on the findings of this study, such teaching interventions could be further refined by increasing the intensiveness of the treatment, for instance through increasing the feedback training instances, increasing the frequency and the amount of feedback to be provided, and adding more self-reflective components to the overall tasks. At the same time, there are indices that suggest that treatments such as the one devised for this study do not suffice to train the competences in question and achieve significant learning effects. Instead, considering to place a stronger focus on TLU-domain exposure in teacher education could have

promising effects on the development of pre-service teachers' profession-related language competences. Of course, any such alterations should be empirically researched accordingly.

7 Research Methodology Sub-Study

This section outlines the methodology of the sub-study conducted to answer RQ #3 and its respective sub-questions. After the description of the research context and research method, I depict the development of the interview guide including the corresponding pilot study. A description of the data collection procedure precedes the conclusion of the subchapter where a methodological reflection including the measures taken to provide valid and reliable findings and ensure adherence to the scientific quality criteria is provided.

7.1. Context and Design Sub-Study

This chapter reports on the small-scale, qualitative, explorative-interpretative sub-study, which aims at providing a complementary exploration of the affordances and challenges of both the PRLCP and the PRLC-R with reference to their systemic relevance, validity and usability. At its core lie the PRCL-R criteria according to which lower-secondary school students voice their perceptions of and evaluate pre-service teachers' profession-related language competences as "field experts" and crucial stakeholders to the application and implementation of the PRLCP and PRLC-R. The aim is neither to produce generalisable findings of the field experts' perceptions nor to make direct comparison between the expert and student judgements. Instead, the sub-study attempts to uncover the heterogeneity and individuality of different stakeholders' needs and perspectives with reference to L2 teacher language competence in the Swiss L2 classroom. Additionally, eliciting student judgements of oral teacher language performances and comparing them with expert ratings serves to gain qualitative insights into the underlying constructs and their practical relevance in L2 education. The aim of this sub-study is thus to investigate the following research questions:

- **RQ #3: How do lower secondary school students perceive and evaluate the linguistic quality and comprehensibility of pre-service English teachers' oral feedbacks in the target language English?**

- RQ #3.1: How do lower secondary school students perceive and evaluate pre-service English teachers' English competence based on oral feedback performances in the target language English?
- RQ #3.2: What (language-specific) aspects of oral feedbacks in the target language English do lower secondary school students perceive as being crucial for ensuring student understanding?
- RQ #3.3: How do lower secondary school students' perceptions of pre-service English teachers' oral feedbacks in the target language English compare to those of trained experts in applied linguistics and English language teaching and learning?

I refrained from forming hypotheses to adhere to qualitative research's most fundamental principle of openness (Caspari et al., 2016; Reimann, 2020a) and instead adopted an explorative, qualitative approach to conduct an in-depth examination and uncover what lies beneath the students' views (Nassaji, 2015). Aside from comprising the research method and instruments so that insights can be gained that could not be hypothesised or expected by the researcher, openness in qualitative research also means that all instruments, methods and convictions or positions can be revised iteratively throughout the entire research process (Reimann, 2020a). This flexibility does not imply that qualitative research can be organised at random but, instead, that the research process needs to be systematic and all steps and stages need to align with the ultimate aim to answer the research questions (Caspari et al., 2016; Reimann, 2020a). The following section outlines the chosen (iterative) research method and the ways in which the above principles of qualitative research have been sought to be met.

7.2. Method

To obtain the necessary data to investigate the research questions, I scrutinised several data collection methods for their appropriateness, including focus group interviews, semi-structured interviews, checklists, or individual rankings of the PRLC-R criteria according to their perceived relevance. I also consulted other ways of eliciting student judgements, such as inductively developing relevant criteria within a small-scale Delphi study. At the same time, it was necessary to test whether the field experts could understand and make sense of the relevant PRLC-R criteria to be able to judge pre-service teachers on their respective language competence in the first place. For this purpose, I conducted a pre-pilot study in June 2019. During a double-lesson, I instructed two lower-secondary school classes to engage with a

simplified version of the PRLC-R to assess the language proficiency of five local celebrities who appeared in a variety of YouTube videos. Even though the students applied most criteria confidently, the criteria *task completion* proved irrelevant, and *cohesion & coherence* too challenging to assess. Therefore, I removed both criteria for the subsequent investigations. However, fully assuring that students would be able to grasp the concepts behind each criterion would require further in-depth examination and exceed both the goal and the scope of this sub-study. After examining the insights from the pre-pilot, I reconsidered the listed methodological approaches and chose to conduct semi-structured interviews that incorporate reframed PRLC-R criteria. Semi-structured interviews are a common form of qualitative interviews that, through their pre-defined structure and impulse-questions, enable the area of interest to be the subject of the conversation (Caspari et al., 2016). They allow for valuable, in-depth insights into individual students' perceptions and a qualitative albeit less rigorous comparison between student and expert judgements. Furthermore, group think bias, recency bias, peer pressure or other social influences that may occur in focus groups can be better mitigated in one-on-one interviews. Interviews also allow the researcher to ask follow-up questions, for which other methods are not as conducive to. As answers are elicited though the interview questions it is important to note that the data cannot be identified as aligning with all principles of qualitative research. It is therefore central for the questions to be formulated as openly as possible to allow the research participants to elaborate sufficiently on their own subjective views (Caspari et al., 2016). In addition, the interview guide needs to contain an introductory part as an icebreaker or warm-up, a main part with all essential questions related to the research subject, and a concluding part with less complex questions as a warm-down (Reimann, 2020a). The following section outlines the interview guide development process in more detail.

7.2.1. Interview Guide

For the development of the interview questions and interview guide I followed an iterative process according to the guidelines of Helfferich (2005), Kvale (2007) and Misoch (2015). As introduced above, the PRLC-R criteria built its foundation. Thus, the overall goal was to create an interview guide that adequately reflects the PRLC-R criteria. To ensure that the field experts achieve an accurate understanding of the underlying constructs of the PRLC-R so that they can apply them for assessment, the below criteria *accuracy*, *vocabulary*, *pronunciation*, *fluency* and *addressee-specificity* were selected as a basis for the interviews:

Wortschatz: Wortwahl	*Sprachliche Korrektheit*	*Aussprache & Betonung*	*Flüssigkeit*	*Adressatenbezug: Lernende*
Sich im gegebenen Kontext mit inhaltlich passender Wortwahl ausdrücken	Sich sprachlich korrekt ausdrücken (Grammatik)	Sich mit korrekter Aussprache und Betonung ausdrücken	Sich flüssig ausdrücken, ohne zu lange oder zu viele Pausen oder Strategien zur Pausenüberbrückung einzusetzen	Sich den Lernenden gegenüber verständlich ausdrücken

Table 30 : Selected PRLC-R assessment criteria for interview guide

After the above criteria were translated into interview questions that seemed appropriate for and accessible to lower secondary school students, further criteria were added such as for instance *overall comprehension.* These additional criteria served to enable participants to access the PRLC-R criteria from both the formal (linguistic) *and* functional (content-specific) view on L2 production. Language teacher educators of the IFDS then reviewed the first draft in a pre-pilot twice until the interview guide was ready to be piloted:

	Leitfrage / Erzählaufforderung	Themen	Präzisierungsfragen
1	**Einstieg: Dein erster Eindruck** *Normalerweise beurteilen Deine Lehrpersonen Deine Leistungen. Heute drehen wir den Spiess um und Du darfst den Lehrpersonen ein Feedback geben und sie beurteilen.*		
1.1	Bitte erzähl mir in einem ersten Schritt: Was ist Dein erster Eindruck dieser Audioaufnahme?	☐ Grund erster Eindruck	- Weshalb hinterlässt die Lehrperson diesen Eindruck auf Dich?
1.2	Was hat die Lehrperson genau gesagt?	☐ Verständnis	- Wie würdest du das Feedback nochmals in deinen eigenen Worten zusammenfassen?
1.3	Gibt es etwas, was du nicht verstanden hast?	☐ Inhalt ☐ Grund	- Was hast Du nicht verstanden? - Weshalb hast Du es nicht verstanden?
1.3	Erzähle mir doch gerne mal, was die Lehrperson ganz allgemein in deinen Augen gut gemacht hat.	☐ Grund	- Weshalb hast Du es so wahrgenommen?
1.4	Was hat die Lehrperson weniger gut gemacht?	☐ Grund	- Weshalb findest Du hat das die Lehrperson weniger gut gemacht?
2	**Sprachgebrauch / Sprachkompetenz** *Jetzt interessiert mich besonders, wie gut Deiner Meinung nach das Englisch dieser Lehrperson ist.*		
2.1	Was hat Dir an der Sprache der Lehrperson gut oder weniger gut gefallen?	☐ Grund	- Weshalb hat Dir dieser Aspekt gefallen? - Weshalb hat Dir dieser Aspekt nicht gefallen?
2.2	Wie gut spricht diese Person Englisch?	☐ Wahrgenommene Kompetenz ☐ Korrektheit	- Wie kompetent wirkt diese Person auf Dich? - Woran erkennst Du, dass die Lehrperson gut/noch nicht so gut im Englisch ist?
2.3	Wie anstrengend war es, die Lehrperson sprachlich zu verstehen?	☐ Sprache	- Wo musstest Du Dich besonders anstrengen, damit Du die Lehrperson verstehen konntest?

Figure 24 : Excerpt interview guide for pilot study

In December 2019, two months after the pre-test of the main intervention study, I piloted the interview questions with three volunteer pupils at a lower secondary school in the canton of St.Gallen during 30-40 minute interviews. For this purpose, I selected four language production samples from the pre-test data pool: two that displayed particularly high, and two that displayed relatively low language proficiency. The research participants could choose between conducting the interview in standard High German or Swiss German so that the language of

instruction could also be piloted. I obtained the participants' legal guardians' formal consent prior to the pilot study and then recorded the interviews with an audio recording device. After the full transcription of all three interviews, I searched and analysed the transcripts for indications regarding the appropriateness and effectiveness of the interview questions. Some of the original interview questions proved to be unclear or repetitive. For example, the opening part of the guide contained questions that prompted the participants to explain how comprehensible they thought the language performance was. Thus, the additional separate middle part on *overall comprehension* mostly led to the same answers, which resulted in the complete removal of this part. I also removed the interview questions that inquired about the usefulness of the feedback in the respective audio recordings. Because the interview was set in an artificial context where a recorded feedback message was presented that was not directed at the participants themselves, they could not relate to it or make an appropriate judgement thereof. Furthermore, I removed the questions that proved to cause difficulty or confusion (e.g., "Fehlen Dir noch Informationen, damit Du weisst, wie Du weiter vorgehen musst?", or "Was wären jetzt Deine nächsten Schritte, wenn Du das Feedback befolgen würdest?"). In addition, I added the following question to the part of the guide that inquires about the participants' judgement of the pre-service teachers' language competence: "Wie gut würde eine Person die Lehrperson verstehen, die im Englisch eher Mühe hat?". By incorporating this question, I intended to gain further insight into the field experts' judgements of the pre-service teachers' language proficiency. I then edited the remaining questions to avoid redundancies and to increase precision. Some questions were shortened to decrease the overall interview length to a maximum of 25-30 minutes. The updated version was evaluated once more by the same IFDS experts, adapted and finalised.

7.3. Research Participants

Five year-9 students (B1, B2, B3, B4 and B5) from the same lower secondary school where I had administered the pilot test a year prior were recruited via their English teachers. I aimed to recruit a) students who are of the same age and school level as the pupil represented in the video-vignette of test task 3, and b) students who differ from one another in terms of their English language skills and academic achievement. The latter was important in order to receive judgements from students who, to a certain limited extent, represent the heterogeneity of English proficiency at a more or less prototypical Eastern Swiss lower secondary school. B1, B2 and B3 visited intermediate level English classes. While B1 and B2's grades in the subject

English were reported as insufficient by their teacher, B3's achievements fell into the more satisfactory middle range. B4 and B5 took the more competitive advanced-level English classes with B4 achieving marks in the middle range and B5 scoring in the upper middle range. Overall, the participants' English proficiency could be located on a CEFR level between A2 and B1. Therefore, participants' English proficiency represented a relatively broad spectrum.

7.4. Research Procedure

I conducted the main data collection in November 2020 after the post-test had been administered and all obtained data had undergone the expert rating process. The following graphic outlines the main steps undertaken for of the qualitative sub-study:

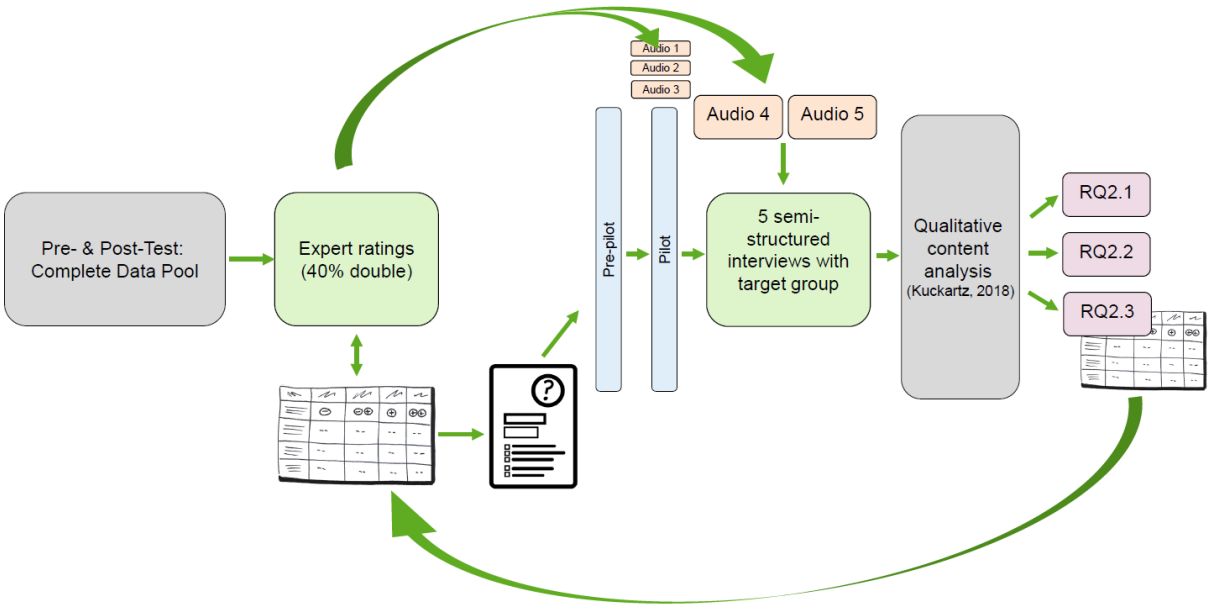


Figure 25 : Outline study design qualitative sub-study

From the complete data pool, I extracted a low and a high proficiency performance sample. The rationale behind only selecting two audio files was to mitigate the cognitive load for both the interviewer and interviewees, prevent clouded judgements and ensure the interview does not exceed 30 minutes. I selected the language production samples based on the following criteria:

- Length and sound quality of the audio files (i.e. language production samples)

- Test task the audio files respond to: both language production samples needed to respond to the same test item to ensure that the research participants could adequately compare between a high- and low-proficiency performance.
- Target group which the test task targets (since the research participants were in their final year of lower secondary school, the test task needed to target same level)
- Expert rating and amount of low ratings per PRLC-R criterion:
    - the most extreme examples (of highest and lowest overall ratings)
    - samples with relatively high interrater agreement
- Additional new global rating by the researcher: overall impression
- Guiding criteria: pronunciation and L1 accent, accuracy and fluency (vocabulary was also important but not the main influencing factor)
- Two separate audio files with very high and low ratings of *addressee-specificity*

Test item 3 from the pre- and post-test proved to be the item that generated the most appropriate audio recordings for the interviews and that best met the above listed criteria:

Richtzeit Aufgabe: ca. 10 min / **Umfang Antwort:** 2-3 min Sprechzeit / **Zielgruppe:** 3. Klasse Oberstufe, Sekundarschule (erweiterte Anforderungen)

Situation:
Sie sie haben im Englischunterricht das Thema «Reisen und kulturelle Unterschiede» intensiv behandelt. Heute führen Sie mit Ihren Schüler*innen kurze Einzelgespräche zum behandelten Thema. Sie wollen die mündlichen Sprachkompetenzen Ihren Schüler*innen formativ beurteilen. Dabei konzentrieren Sie sich auf die Kategorien *Inhalt* und *sprachliche Korrektheit* auf dem Ihnen ausgeteilten Beurteilungsraster.

Im Einzelgespräch haben Sie Nathalie die Frage gestellt, in welchen Reise-Situationen man sich als Tourist besonders angemessen verhalten sollte.

Machen Sie sich mit dem Ihnen ausgeteilten Beurteilungsraster und der darunter stehenden Aufgabe vertraut. Schauen Sie sich danach das Video mit Nathalies Antwort an und machen Sie sich Notizen, auf welche Sie sich bei der anschliessenden Rückmeldung stützen.

Aufgabe:

Geben Sie Nathalie auf **Englisch** eine kurze Rückmeldung zu ihrer Fähigkeit, sich in dieser Situation **angemessen** ***(«Inhalt»)*** **und sprachlich korrekt** ***(«sprachliche Korrektheit»)*** auszudrücken.

Stützen Sie sich bei Ihrer Rückmeldung auf das Beurteilungsraster und Ihre Notizen.

Nehmen Sie Ihre Rückmeldung **auf Englisch** auf. Beachten Sie die angegebene Zeitvorgabe sowie das Niveau der Klasse. (**Sprechzeit** 2 - 3 min).

Figure 26 : Test item (task 3 of pre-/post-test) selected for the semi-structured interviews

Kategorie	☹	😐	☺	Kommentare
Sprachliche Korrektheit	❑ Sie/er macht häufig Fehler.	❑ Sie/er macht manchmal Fehler.	❑ Sie/er macht nur sehr selten oder gar nie Fehler.	
Inhalt	❑ Die Wortmeldung ist inhaltlich unpassend.	❑ Die Wortmeldung ist inhaltlich grundsätzlich passend.	❑ Die Wortmeldung ist inhaltlich treffend.	

Table 31 : Additional test task resources of test item 3

Analogous to the pilot test, the students' legal guardians' permissions were obtained and the interviews audio-recorded. All interviews apart from one were conducted and transcribed in standard High German. Before each interview, I introduced each participant to the overall research project as well as to test item 3 where the particular language production samples were

taken from. They were then familiarised with the task they were about to complete and introduced to the concept of the role-reversal they were going to engage in. I emphasised that the participants could interrupt and ask questions at any time, request to re-listen to the audio files as often as they needed, and terminate the interview at any point with no explanation. With this detailed contextualisation of the task and roles, I sought to foster an environment where I could elicit judgements from the participants that were as objective as possible. Afterwards, I fully transcribed all interviews verbatim in MS Word with close temporal proximity to the interview dates and adhered to the transcription conventions of Dresing and Pehl (2013) and Kuckartz et al., (2008). To ensure uniform spelling throughout, I followed the guidelines by Dresing and Pehl (2013) (e.g., capitalising phrases or expressions with strong emphasis). I removed all names and references to persons to anonymise the data. However, since there were only five research participants and I had conducted, transcribed and analysed the interviews myself, the participants remained recognisable both through their voices in the audio recordings as well as through the statements they made during the interviews. To analyse the interviews, I conducted a structured content analysis according to Kuckartz (2018), which will be described in chapter 8.

8
Data Analyses and Results Sub-Study

The present sub-study is interpretative, qualitative and explorative in nature and contains a strong focus on the practical implementation of the PRLCP and PRLC-R in the authentic L2 teaching and learning context. This chapter outlines the qualitative content analysis of the interview transcripts (chapter 8.1), presents the findings with regard to lower-secondary school students' perceptions of pre-service English teachers' spoken language performance on the example of oral feedback provision, and outlines the established answers to RQ #3, RQ #3.1, RQ #3.2 and RQ #3.3 (chapter 8.2).

8.1. Qualitative Content Analysis Sub-Study

To analyse the interview transcripts, Kuckartz' (2018) hermeneutic approach to structured qualitative content analysis was selected. This method is characterised through the systematic allocation of text passages to a designated system of categories that can be developed deductively, inductively, or through a combination of both. Kuckartz's approach is a frequently applied and widely recognised content- and theme-oriented method for qualitatively analysing texts (Schreier, 2014). At its core lie the identification and conceptualisation of aspects derived from the textual material and the subsequent systematic description of the material based on the identified aspects. These simultaneously build the structure of the system of categories. Finally, the identified aspects are subsumed into overarching themes, which then again shape and explain the individual categories of the category system (Schreier, 2014). The aim of this sub-study is not to achieve alignment in the categorisation of themes, but to analyse emerging themes and interpret them accordingly. Kuckartz's (2018) qualitative content analysis approach enables an analysis that allows for close proximity of the researcher to the available textual data throughout the entire process. It is therefore in line with the research aims of RQ #3. The following section reports step-by-step on the analysis I conducted of the present interview texts.

8.1.1. Phase 1: Initiating Textual Analysis

The initiating textual analysis constitutes the first phase of qualitative content analysis. It involves the in-depth, sequential and systematic study and annotation of the entire textual linguistic material (Kuckartz, 2018). These annotations are called *memos*[27] and form an integral part of the study of the text (ibid.). The purpose of this phase is to gain an overall understanding of all texts and their subjective meaning and value with reference to the research questions (ibid.). I implemented this step by highlighting important passages of each transcript and annotating them with memos using the comments function in MS Word:

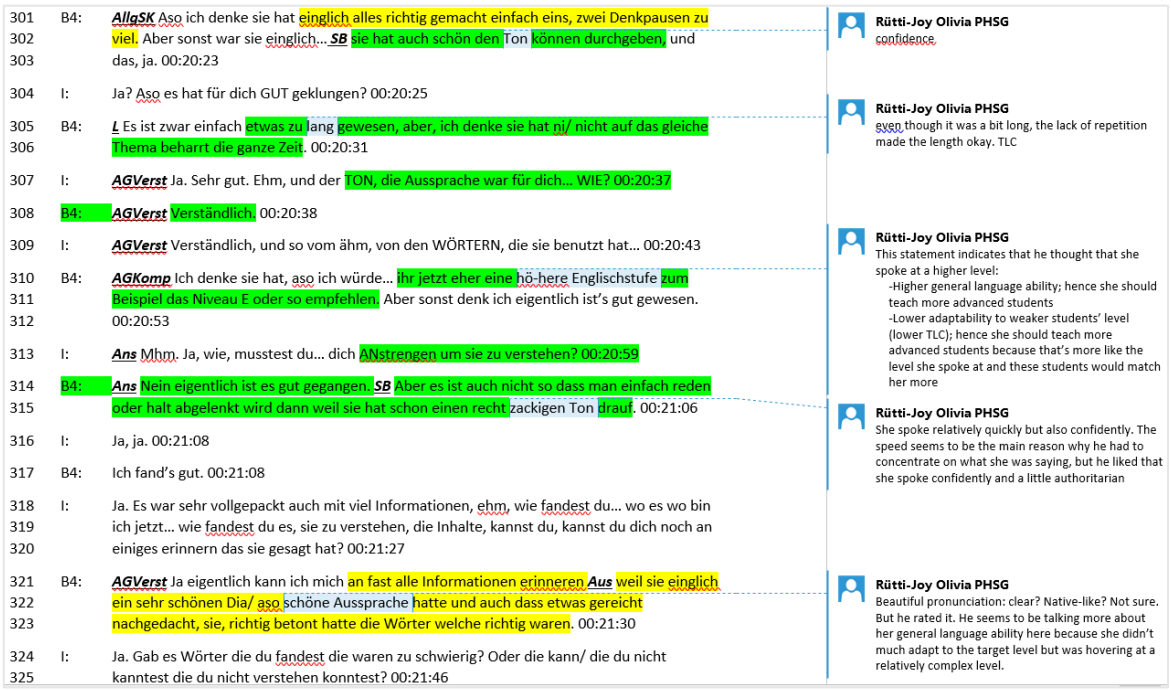


Figure 27 : Example of annotation with memos during initiating textual analysis.

Subsequently, I composed a case summary[28] for each research participant. This summary is fact-oriented, close to the text and free of interpretations. Assumptions that cannot explicitly be proven through text passages need to be explicitly indicated (ibid.). Designating a label to each case that characterises the specifics of the particular interview or participant (in terms of a "motto") completes the summaries. Kuckartz (2018) assigns case summaries a four-fold purpose, arguing that case summaries

[27] Memos are collections of the researcher's thoughts, ideas, assumptions and hypotheses that arise during the data analysis and form an integral constituent of the entire research process (Kuckartz, 2018, p. 58).
[28] Case summaries are systematic summaries of the characteristics of each particular case with reference to the research questions (Kuckartz, 2018, p. 58).

- allow larger research teams to access the data through an overview of the essential content even if not all members systematically worked through each text (team-aspect).
- provide a useful basis for developing case overview tables (comparative-aspect).
- allow the researcher to approach the data with a more focused, analytical stance to recognise divergences or convergences between different cases (aspect of analytical differentiation).
- can contribute to the generation of hypotheses and categories.

Table 32 below presents the case summaries for each research participant:

Interview B1: The proponent of self-confidence	
Formal and structural characteristics: • 25:57 minutes • Interview in Standard High German • Mostly speaks in first person singular, when he criticises he tends to speak in the 3rd person singular • Speaks relatively fast and seems like he wants to get it over with *Background about participant:* • Lower level class • Experience learning English: 6½ years • Marks: insufficient and substantially lower than a 4 (doesn't put in any work, good speaking skills, makes almost no progress)	*Case summary:* Likes if a teacher displays self-confidence (motivation / enthusiasm / alertness / presence). Finds the message and the teacher convincing if they come across as self-assured and self-confident; If it is evident that the teacher puts in effort for their students, he finds the teacher and his/her language skills convincing. Links a "native-sounding accent" and a convincing tone of voice to advanced language proficiency; someone who is confident and speaks as confidently as (and sounds like) a native speaker would is convincing to him. Considers clarity, fluency and speaking at the right volume as aspects that contribute to student understanding and as indicators of proficiency. Thinks long and repetitive monologues or lectures are not very conducive to learning. Considers frequent hesitations, the frequent use of fillers and frequent and long pauses to find the right expression as indicators of low proficiency. Says that hesitations, pauses, fillers and low volume (linked to low self-confidence? *Interpretation*) make it difficult for the listeners to follow the message. Hesitations are distractors; a "native-sounding accent" and confidence are attractors. He values when it is evident that a person has a lot of experience with a language and has perhaps even lived in a country in which the target language is spoken. Distinguishes between "basic" and "high" language skills ("basic English and educated English"), which he recognises based on the complexity of the vocabulary used.

	Speaking at a slow(er) pace and taking enough time for weaker learners is conducive to student understanding. Found it difficult to judge the teachers' language proficiency based on formal language criteria because he considers his English skills not to be advanced enough to do so.
Interview B2: The advocate of a friendly-sounding tone	
Formal and structural characteristics: • 21:16 minutes • Interview in Standard High German • Tendency towards short answers, does not elaborate much. Very much a ping-pong game between the interviewer and the interviewee *Background about participant*: • Lower level class • Experience learning English: 6½ years • Marks: only just sufficient reaching a 4	*Case summary*: A friendly tone and nicely embedded feedback is important to feel comfortable: a kind and friendly teacher is valued. To speak clearly, at an appropriate and well-intelligible volume and at a moderate pace is important to her and for her understanding. If the teacher mumbles and speaks fast, she finds it difficult to follow. Notices pronunciation and vocabulary when it comes to judging language ability. Hesitation, disfluency and self-correction is noticed and seems to be connected to lower language ability (*interpretation*), however pauses and hesitations do not bother her. Clarity of speech is more important for ensuring understanding than pace: if clarity is given, the pace may also be faster and still be intelligible. Has an awareness of people's multilingual/plurilingual repertoire: using an additional language to make connections could help when explaining something to weaker learners in English. Recognises that teachers could use more low-frequency words if they wanted to, but that they make a conscious decision to adapt their speech to the level of their students. The precision and complexity of someone's vocabulary are an indication of higher language ability to her.
Interview B3: The language-aware	
Formal and structural characteristics: • 33:30 minutes • Interview in Swiss German as requested • Elaborated on her answers • Most of her statements are in the first person singular; however when she criticises she tends to use the second person singular ("you" as opposed to "I" for distancing herself from	*Case summary:* There is an awareness of teacher language proficiency and that it does not equal general language ability: her notions of fluency, speed, volume and complexity of vocabulary indicate that there is a difference between high general language proficiency and high teacher language proficiency (interpretation). She sees that the teacher spoke slowly and fluently (speed and fluency are mingled up here → *interpretation*) for the purpose of teacher-student communication, however that the teacher's ability would include higher fluency too.

the critique) to explain where problems lie *Background about participant:* • Lower level class • Experience learning English: 6½ years • Marks: middle range (4.5 – 5), studies at contributes in class	She considers there to be a need for breaking down complex content into manageable chunks to ensure student understanding. Pronunciation and variety / accent: an inclination towards native-speakerism (interpretation). If an "accent" corresponds with her notion of beauty / aestheticism in a language, she loves listening to it. If it does not correspond to it, she tends to get bored and to not to want to listen anymore. Considers high fluency, accurate and «really good pronunciation» and high accuracy as indicators of high language proficiency. Pronunciation as a criterion seems to be understood as an umbrella term, which also encompasses fluency and loudness (*interpretation*). Disfluency and inappropriate volume equals bad or inappropriate pronunciation. She considers positive reinforcement as important. Clarity and appropriate loudness are necessary components for ensuring understanding. Fillers, pauses and hesitations are noticed and considered as markers of disfluency. Disfluency, hesitations and fillers disrupt or impede understanding. Knows that a predictor for fluency is language ability, which to her means knowing how to structure sentences and applying different language skills and language knowledge appropriately.
Interview B4: The sub-challenged who needs boundaries	
Formal and structural characteristics: • 32:06 minutes • Interview in Standard High German • Chatty and generously elaborated on his answers. *Background about participant:* • Advanced level class • Experience learning English: 6½ years • Marks: middle range	*Case summary:* Appreciates when teachers make an effort for their students, both in terms of adapting the complexity of their speech to the target audience and in terms of providing good and helpful feedback. Thinks that by adapting the articulation rate and complexity of the vocabulary to the target audience, understanding can be ensured. Prefers when teachers get to the point and explain only the most important points concisely and precisely. To him, this contributes to ensuring understanding, especially also for weaker students. While he generally sees repetition as a distractor, in the case of weaker learners repetition could contribute to their understanding. Thinks that a confident manner of speech, appropriate volume, confident tone and evident enthusiasm for teaching and the students are important skills a teacher needs to have. These skills enable classroom management and ensure understanding (because it is only then that students would actually listen).

	Thinks that disfluency can disrupt the flow and contribute to losing the students' attention. A teacher with very proficient command of English (which, for him, is noticeable by more low-frequency vocabulary) had better teach more advanced students because it would be a better match for them. Notices repetitions, new starts and hesitations, sees them as an indication of disfluency and hence of lower language ability (last point: *interpretation*).
Interview B5: The ambitiously curious	
Formal and structural characteristics: • 29:29 minutes long • Interview conducted in Standard High German • Chatty and elaborated on her answers. *Background about participant:* • Advanced level class • Experience learning English: 6 years • Non-native German speaker • Marks: upper middle range	*Case summary:* Appreciates a positive tone but sees a need in teachers to be critical of students' performance in order to help students improve (constructive feedback is important). Constructive criticism with clear statements on what needs to be improved, and why and how it needs to be improved, is important to her. Appreciates precise and clear pronunciation. Likes when a teacher is confident and not nervous, as the latter contributes to clearer speech and ensures understanding. Finds it important that the teacher encourages the students, shows enthusiasm for teaching and displays interest in the students. Notices hesitations, pauses, fillers and pronunciation as indicators of general language ability. Fillers and hesitations are seen as distractors that impede understanding. To be *interpreted* against the background that this participant is a non-native German speaker and might have a different motivation to learn English than her peers. Clear speech and speaking at an appropriate volume contribute to better understanding.

Table 32 : Case summaries of qualitative, semi-structured interviews

Once I had completed the case summaries for all research participants, I could initiate the next phase of the qualitative content analysis: building the categories.

8.1.2. Phase 2: Construction of Thematic Main Categories

When conducting qualitative content analysis, working with and on a textual category system and employing the categories in a reflected and informed manner is of central importance (Kuckartz, 2018, p. 83). Hence, category-formation takes on a crucial role within the overall qualitative content analysis. The literature on research methodology distinguishes between two

opposite polarities of category formation, which are connected through a wide spectrum of mixed-methods-approaches (ibid.). The deductive, theory-based category formation method (i.e. *a-priori-category-formation*) is characterised through deriving the categories from an already existing system, be this a theory, hypothesis, an interview guide or other pre-existing material (Kuckartz, 2018). The inductive, databased method, on the other hand, includes deriving categories directly from the available textual material (ibid.). Both approaches can never be entirely objective. They both involve an active construction process and active allocation of the existing material to categories, which is always dependent on and influenced by factors such as the researcher's prior knowledge or language competence in the language in which the categories are built (Kuckartz, 2018, p. 72). It is rare that research projects employ a category-formation-approach that is purely inductive or deductive. Rather, multi-stage mixed methods approaches tend to be much more common, especially when instruments such as interview guides are used for data collection. This approach is also referred to as *deductive-inductive category formation* - the approach I chose for the present sub-study. I derived the categories from the interview guide in a first step and subsequently further developed them through building sub-categories or entirely new categories based on the empirical material itself. One of the biggest challenges when forming categories is to ensure that their descriptions are precise enough so that any overlap can be avoided (Kuckartz, 2018, p. 67). Furthermore, it is crucial that the category system is complete and does not lack any important aspects that need to be considered when analysing the data. Hence, the categories need to be clear-cut, disjoint and exhaustive to meet the standards of the scientific quality criteria for coding (Kuckartz, 2018, p. 67). At the same time, it is important that the categories plausibly and meaningfully relate to one another and build a coherent and exhaustive system (Kuckartz, 2018, p. 71). To build the categories, I combined Kuckartz' guidelines for inductive and deductive category formation as outlined below (Kuckartz, 2018):

1) Goal-definition of the category-formation based on the underlying research questions

This sub-study seeks to investigate a) how students perceive and evaluate pre-service teachers' English proficiency and b) uncover what (language-specific) aspects students consider as crucial for ensuring their understanding. This includes finding out what students notice about pre-service teachers' spoken L2 proficiency when asked to evaluate oral feedbacks, and what language-related characteristics the students consider important for an English teacher to have in order to conduct good English lessons. It is not the goal to categorise the research participants into types. Rather, the sub-study's focal point lies in uncovering individual perceptions and

preferences in form of important themes, and analysing those in light of the PRLCP and PRLC-R. As the interview guide is based on the PRLC-R, the initial deductive categories were derived from the interview questions themselves. This sub-study has a strong focus on the application of the rubric and the profiles in the field. Therefore, by answering the RQ #3.1-#2.3, I aimed to gain additional insights and shed light on complementary perspectives to enrich the main-study.

2) Determination of the type of category and level of abstraction

Given the explorative nature, the corresponding research questions, and the interest in uncovering individual preferences and perceptions of the research participants, thematic codes are the most appropriate type of category for this analysis. A thematic code describes a specific theme, argument or mental figure (Kuckartz, 2018). In qualitative content analysis, the textual passages, which contain information on the different themes, are assigned to the respective thematic code (Kuckartz, 2018, p. 34). The codenames are to be of low abstraction and the inductive categories close to the primary text. This is due to the interest in accessing and representing the research participants' perceptions as accurately as possible. In contrast, to establish coherence between the interview questions, the interview guide and the participants' answers, the deductively derived categories may be more abstract and closer to the classifications used in the primary material (PRLCP and PRLC-R).

3) Familiarisation with the data and determination of the type of coding unit

This familiarisation process took place during the initiating textual analysis and annotation of the textual material with memos as described above (chapter 8.1.1). The coding unit refers to a single element that initiates an allocation to a category. A coding unit thus needs to be defined so that it is still meaningful if removed from its context (Kuckartz, 2018, p. 104). In the case of this sub-study, a coding unit is defined to be theme- and content-related and is henceforth referred to as thematic coding unit. Here, a coding unit corresponds to one clear meaning component (a semantic unit) in the text; i.e. one unit of meaning that is in itself coherent and self-contained (see appendix G). A unit of meaning in the text refers to a particular theme whose length can vary from one phrase to several turns in the interviewer-interviewee interaction, as illustrated in the following two examples:

Example 1: *B1: es war halt, sehr LEISE*

Example 2: *B1: Das Tempo war langsam. Nicht ZU langsam aber ein bisschen langsamer als das in als das perfekte Mitte. 00:21:58*

I: Ja. Also für dich persönlich wär es ZU langsam? 00:22:01

B1: Ein bisschen langsam, ja wahrscheinlich. 00:22:04

I: Würdest du dich langweilen 00:22:06

B1: Aso für mich, ich hätt es gern mittelmässig. Nicht zu schnell nicht zu langsam. 00:22:08

In contrast to common misconceptions, a single thematic coding unit can stand for more than one category simultaneously because textual passages and sentences can contain several themes (Kuckartz, 2018, p. 41, 102-103).

4) Sequential and systematic processing of the textual material, inductive category formation, and allocation of pre-existing categories or formation of additional categories

Category formation is an iterative and constructive process, which demands constant reflection. During its initial stage, it is important to start forming categories with no overt restrictions or limitations to ensure that no important information is lost. This openness does not merely allow but require the researcher to continually ask questions with reference to the research questions, the study aims, the potential pre-existing categories and the creation of new ones (Kuckartz, 2018, p. 84). Depending on the data volume, the process of inductive category formation can be undertaken on a subset of the texts (Kuckartz, 2018, p. 84). As the present data pool contains five interview transcripts, I built the inductive categories in a single-phase-procedure by working through all the available textual material sequentially. Even though it is recommended to choose the sequence of textual material randomly to avoid potential bias or distortions, given the limited amount of available data, I processed the interviews according to the chronology of when they had been conducted. In a first step, I worked through all interview transcripts by allocating to textual passages the deductive categories created from the interview guide. Simultaneously, I derived new inductive categories directly from the interview transcripts. Relevant passages that exemplify specific categories were extracted from the textual material to supplement the coding frame and act as anchor examples. Next, I revisited the transcripts to further develop and differentiate the existing categories and to form additional categories.

5) Systematisation and organisation of coding frame

This step is usually undertaken when a certain level of saturation of categories has been reached or when the amount of categories becomes too large to keep a systematic overview. The systematic organisation of the coding frame inherently involves tidying up the categories,

subsuming those who are closely related, organising them into main categories and sub-codes and organising the collection into a logical system. A coding frame can take many forms of organisation, and for the purpose of this study, I chose to organise them according to different perspectives on feedback: (1) Feedback als sprachliche Produktion, (2) Feedback aus Sicht der Rezeption: Feedback als Interpretation; (3) Feedback als pädagogisches Werkzeug; (4) Feedback als soziales Artefakt zur interaktionalen Mediation von Wissen, Verständnis und Lernfortschritt: Mediationskompetenz; and (5) Forschungsteilnehmende in der Rolle als Beurteilende. Like the previous steps, the systematisation and organisation was iterative, and involved moving back and forth between the interview transcripts and the coding frame.

6) Finalisation of coding frame

Once saturation was reached, I examined the coding frame for its adherence to the scientific quality criteria, ensuring that the categories were transparent, succinct, plausible, exhaustive, comprehensible, and clear-cut (Kuckartz, 2018, p. 85). After this revision process, the coding frame was finalised and each category annotated with a precise definition and anchor example (i.e. linguistic example extracted from the text) where they were still missing. It is important to note that this finalisation does not equal a final, conclusive completion – the coding frame remains alterable and adaptable to new discoveries that may emerge during the coding of the remaining material (Kuckartz, 2018, p. 86).

8.1.3. Phase 3-6: Coding

Once the coding frame was finalised, I started with the first round of coding (phase 3) by sequentially coding all textual material. For this, I established the following coding rules:

- At least one phrase constitutes one coding unit, given that it clearly corresponds to one semantic unit (i.e. one meaning component).
- One meaning component can subsume one utterance up to several turns.
- A coding unit contains enough meaningful material for it to make sense if it was removed from the text.

The coding was conducted using the software QCAmap[29] (ASQ, 2021). After completing the first round of coding, the subsequent steps involved further differentiating the existing categories, which resulted in the emergence of more sub-categories. By adding the new sub-codes to the existing categories, the coding frame was finalised as follows (see Appendix G):

Thematic main category	**Subcategory**
Feedback als sprachliche Produktion	
Sprachkompetenz	Sprachkompetenz: Allgemein
	Erfahrung
	Korrektheit
	Wortschatz
	Flüssigkeit
	Sprechgeschwindigkeit[30]
	Native-Speakerism
Feedback aus Sicht der Rezeption: Feedback als Interpretation	
Verständnis	Verständnis: Allgemein
	Anstrengung
	Verständlichkeit für schwache Lernende
Verständliche Aussprache	Verständliche Aussprache: Allgemein
	Varietät
Selbstbewusstsein	-
Auditive Sprachwahrnehmung: Psychophonetik / Psychoakustik[31]	Sprechmelodie / Tonfall als parasprachliche Funktion der Prosodie (Ebene der A-Prosodie[32])
	Subjektiv wahrgenommene Tonhöhe (pitch): Mel
	Subjektiv wahrgenommene Lautheit (loudness): Sone
Feedback als pädagogisches Werkzeug	
Pädagogisches Wissen (PK)	Pädagogisches Wissen: Feedback
	Erklären: Methode
Adressatenbezug	Adressatengerechtigkeit: Allgemein
	Adessatengerechter Wortschatz

[29] QCAmap is an interactive web application and qualitative data analysis software operated by the Association for Supporting Qualitative Research ASQ. See https://www.qcamap.org or Fenzl and Mayring (2017) for more information.

[30] This category refers to the articulation rate of a speech production. Articulation rate is defined here as the "pruned" syllables per second: the total number of syllables produced excluding dysfluencies (e.g., filled pauses, repetitions, self-corrections, false starts), calculated over the total duration of the speech sample).

[31] Psychoacoustics refers to the study of sound perception. It includes aspects such as loudness (Lautheit) and pitch (Tonhöhe). The subjective perception is entitled as *sone*, the subjective perception of pitch is called *mel,* and the loudness at which a person perceives a sound is called *phon* (cf. Zwicker & Fastl, 2007)

[32] A-prosody is a function of sound production that can be deliberately controlled by the speaker. Parameter of a-prosody include, among others, intonation, pauses, and a change in loudness. Such functions help a speaker to convey the meaning and intention of a speech production and reduce syntactical and lexical ambiguities (cf. Tillmann, & Mansell, 1980).

	Adressatengerechte Komplexität Inhalt adressatengerecht portionieren
Verbesserungstipps	Feedback Sprache
Vertrautheit Feedback	-
Feedback als soziales Artefakt zur interaktionalen Mediation von Wissen, Verständnis und Lernfortschritt: Mediationskompetenz	
Interaktionskompetenz	-
Engagement	Engagement: Allgemein Mühe geben Motivieren
Überzeugende Stimme und Ausdruck	-
Inhaltliche Redundanz Feedback	-
Forschungsteilnehmende in der Rolle als Beurteilende	
Schwierigkeit Beurteilen	
Spass, die Lehrperson zu beurteilen	

Table 33 : Condensed version of the finalised coding frame

Finally, I used the finalised coding frame to recode all textual material a second time. The following excerpt represents an example of a code, its categorisation and the corresponding anchor example, i.e. an illustrative linguistic sample from the text:

Abbreviation	Thematic main category	Sub-category	Definition	Examples
RQ1-5: SKFL	Sprach-kompetenz (SK)	Flüssigkeit	Aussagen zur Sprechflüssigkeit, Zögern und Sprech-pausen, sowie zum Einsatz von Füll-wörtern und weiteren Strategien zur Pausenüberbrückung.	es war… MEHR als mittelmässig , es war nicht SEHR flüssig, und auch nicht GAR NICHT flüssig, es war MEHR als mittel-mässig, aso es… FEHLT noch ein bisschen ah si/ wa/ GANZ ein bisschen unsicher

Table 34 : Example code with anchor example

The completion of this process initiated the next phase of the analysis: creating case-specific, thematic summaries. This step is often also referred to as "framework analysis" (Kuckartz, 2018, p. 111) and includes the creation of a matrix through condensing the primary material by systematically and analytically summarising the coded passages. These case-specific thematic summaries proved to be a valuable resource for the subsequent analysis of the data, as described in the subsequent section below where the results are introduced.

8.2. Results Sub-Study

The final phase (phase 7) encompassed the data analysis, first along the main categories and subcategories, and then between the main categories and subcategories. For this step, I drew on the memos, case summaries, textual case summaries and coded passages while collecting and synthesising emerging themes and seeking an answer to the research questions. This section reports on the findings resulting from the content analysis of the interviews. The overarching research question states the following:

> RQ #3: How do lower secondary school students perceive and evaluate the linguistic quality and comprehensibility of pre-service English teachers' oral feedbacks in the target language English?

This sub-study can be considered a derivative of a small-scale rater cognition study, whereby the target population is drawn upon as field experts (i.e. novice raters) to identify aspects they consider valuable and criteria they recognise as important when it comes to L2 teachers' language proficiency. This research is concerned with uncovering the features the target group notices when listening to (pre-service) teachers' spoken language performances and focus on when evaluating said language productions.

8.2.1. A Note on Audio-Speech-Samples

As the research participants' perceptions of pre-service English teachers' language productions were restricted to audio recordings, visual and contextual information as well as slightly more peripheral aspects of extralinguistic, paralinguistic and non-linguistic cues (e.g., gestures, eye-contact, facial expressions such as smiling or frowning, posture, visual markers of confidence, etc.) was lacking. This could potentially affect the validity of the pupils' judgements, as an ample body of research on listening perceptions of oral performances suggests that visual information presents an important aspect that listeners rely on to grasp a spoken text (e.g., Burgoon, Guerrero, & Floyd, 2016; Raffler-Engel, 1980). For example, investigations on test-takers' listening comprehension indicate that video-based test items better facilitate test takers' understanding than audio-based tasks – precisely because visual and contextual information is available in the former (e.g., Wagner, 2008, 2010). Nakatsuhara et al.'s 2020 study supports such findings, as their research indicated that reliably rating audio-recorded speech productions was considerably more difficult than judging live and video-recorded performances due to the

compromised access to essential additional information in the latter. Even though contextual visual information delivers a more comprehensive picture of someone's speaking and communicative ability, the degree to which assessors should consider non-linguistic features, however, remains disputed (Nakatsuhara et al., 2020). While video-based test tasks may contribute to higher face validity, there are concerns that the visual information may distract test takers as it may increase the cognitive load during an already highly cognitively demanding task (Bejar et al., 2000). As the research participants of this sub-study were asked to complete the cognitively challenging task of judging people whose English language proficiency was substantially more advanced than their own, the effect of the missing visual information on the validity of their judgement can be interpreted along both lines of argumentation. Only hearing the language productions meant that pupils could exclusively focus on their own auditory perceptions, and do so perhaps more scrutinisingly than they would in a face-to-face, visually enriched communicative context. These circumstances may have skewed the results as the research participants may have tended to over-focus on certain auditory aspects. On the other hand, the high cognitive load while performing the task may have been compromised slightly by removing visual information. Still, the difficulty of the task may have also been overwhelming to the participants, resulting in them randomly guessing an answer. Therefore, the results of this sub-study need to be interpreted with caution. However, with the sub-study being designed to adhere to scientific quality criteria, being explorative and rejecting attempts to generalise findings, these limitations do not render the results invalid.

8.2.2. Findings RQ #3.1

Qualitative content analysis revealed a number of deductively and inductively built main and sub themes that lower secondary school students noticed when assessing the language productions. The following section presents the findings to RQ #3.1:

> RQ #3.1: How do lower secondary school students perceive and evaluate pre-service English teachers' English competence based on oral feedback performances in the target language English?

In particular, the research participants identified six overall indicators for high, and two main indicators for low language proficiency respectively:

Indicators for high language proficiency	Indicators for low language proficiency
"Good pronunciation"	No explicit mention of what "bad pronunciation" would be, except for unintelligibility of certain varieties (e.g., Indian English)
High fluency	Disfluency including hesitations, repetitions, fillers, false starts, and pauses linked to searching for vocabulary
Use of low-frequency words	
Ability to articulate a message concisely	
High articulation rate	
Error-free language production (accuracy)	

Table 35 : Indicators for language proficiency

One of the most frequently mentioned categories with reference to pre-service teachers' English proficiency was pronunciation. Indeed, all research participants noticed specific pronunciation features and stated that "good pronunciation" is a clear indicator of someone's L2 proficiency:

> *I: Ehm, woran erkennst du, denkst du, dass diese Person GUT im Englisch ist?*
>
> *B1: Ehm, also a/ an der Aus/ AUSsprache, merkt man das, und, wenn sie, halt, eine überzeugende Stimme dazu hat. Dass… daran merkt ma/ man, dass, s/ s/hat dass sie's kann. (Interview transcript B1, line 84-87)*

There seems to be consensus among the participants that "good" pronunciation is a key indicator of language proficiency; what constitutes "good" or "bad" pronunciation, however, was not elaborated on directly. Yet, occasional references to a speaker's accent or variety of English lead to the assumption that "accent" (i.e. variety) may constitute a factor that shapes the perception of "good" pronunciation. These references were inconsistent when it came to connecting them to pupils' perceived language proficiency: While some participants associated a native-sounding variety (e.g., British English) with high language proficiency – which could be interpreted as a potential underlying tendency towards *native-speakerism*[33] – others found such varieties acoustically pleasing and agreeable to listen to. With reference to *native-speakerism*, two participants indicated that the language proficiency of a native speaker is to be

[33] In L2 research, *native-speakerism* manifests the ideology that native-like language proficiency is equated to effective teaching. Holliday (2005) describes *native-speakerism* "as a social and theoretical position which asserts that so-called "native speakers" are the best models and teachers of English because they represent a "Western culture" from which spring the ideals both of English and of the methodology for teaching it" (p. 6). This approach originates in the assumption that "being a 'native' member of a language community fosters cultural and linguistic knowledge which can translate into both the content and processes of classroom teaching" (Freeman, 2017, p. 33).

aspired to and equated with high language proficiency. They recognised this by the teachers' pronunciation and indicated a preference towards "native-sounding" speech. A teacher who speaks like a native speaker was thus considered a good role model and was trusted to teach them "what it's really like" in the "real world":

> *B1: Ich finde das GUT, dass man halt, wenn man eine Sprache… unterrichten will, dass man auch so zeigt wie ist es wirklich ist. Was bringt es mir wenn ich in der, wenn ich die Deutsch, das Deutsch-Englisch in England anwende. (Interview transcript B1, line 113-115)*

In this statement, B1 seems assume that the variety of English spoken in England (or any other English-speaking country) is automatically the "right" kind of English. He thus indicates a lack of awareness of the multitude of L2 speakers and multiculturalist society of today's globalised world. Most participants however did not show any explicit inclination towards native speakers. For participant B3, for example, a non-native-sounding variety did not automatically indicate low(er) L2 proficiency. With reference to *pronunciation*, another sub-category emerged from the data: self-confidence and a convincing (audible) appearance. Three students explicitly mentioned that they liked a self-confident-sounding tone of voice or any other signs that indicate self-confidence and self-assurance. A self-assured-sounding voice led two students to believe that the perceived self-confidence was rooted in a language biography that indicates ample experience with the L2, which was consequently seen as an indicator for high L2 proficiency. A timid-sounding or quiet tone of voice, in contrast, was perceived as an indicator for low self-esteem and (language) insecurities, which was associated with low(er) language proficiency. The second most referenced category the research participants linked to language proficiency was *fluency*. All field experts mentioned either similar or identical aspects to be markers of high and low fluency respectively. For example, students agreed that high fluency indicates high language proficiency; one even mentioned that disfluency was caused by language insecurities, which in turn was considered an indicator of low(er) language proficiency. Similarly, high fluency was automatically connected with high language proficiency. At times, high articulation rate was also seen as a marker of fluency and hence, in their interpretation, as a marker of high proficiency. It is worth mentioning that in this particular case, articulation rate seemed to be equated with fluency rather than understood as a distinctly different criterion. Unlike with *pronunciation* or *fluency*, (grammatical and lexical) *accuracy* seemed to be categories that the research participants particularly struggled with. From their

point of view, "error-free" language use (be this in term of lexis and / or grammar) was interpreted as being an indicator of advanced L2 proficiency. Considering that the participants' English proficiency corresponds to roughly a CEFR level A2-B1, it is not surprising that this conviction was more of an idea than an aspect they could identify directly from listening to a speech sample (see chapter 7.3). Indeed, they self-reported finding it difficult, showed insecurities while assessing lexical accuracy and, while doing so, appeared to make random guesses as presented in the example below:

I: wie fandest du ha/ war der Wortschatz von dieser Person?

B2: Aso sie hat paar GUTE Wörter eingesetzt, aber… sonst… ni/ nicht so… ja. (Interview transcript B2, line 213-215)

Still, two pupils connected an obvious and recurrent use of high-frequency words with low language proficiency:

B4: Ich würde eine sechs [von zehn] geben

I: Eine sechs? Was sind die Gründe für diese Note?

B4: Aso ich denke sie hat vielmal das Wort «also» oder, halt, die gleichen Wörter eigentlich benutzt. (Interview transcript B4, line 369-372)

Comments on grammatical accuracy were close to non-existent; only one pupil peripherally made a reference while summarising what she recognised as indicators of language proficiency:

B3: I dengg dasch, da ghöri gad vo usch vo i weiss nöd, jo. Wenn öpper so schnell Englisch redt oder so, jo halt eifach so, halt so richtig gueti Uusproch het und n/ kein einzige Fehler macht, DENN (Interview transcript B3, line 84-86)

Both judging from the lack of mention and from pupils' self-reported difficulty it seems evident that (grammatical and lexical) *accuracy* as an assessment criterion posed a particular challenge for the students. It seemed to be both a struggle to evaluate *and* to recognise it.

8.2.3. Findings RQ #3.2

When attempting to answer RQ #3.2, qualitative content analysis revealed a number of aspects that students noticed that they considered as enabling or impeding understanding.

RQ #3.2: What (language-specific) aspects of oral feedbacks in the target language English do lower secondary school students perceive as being crucial for ensuring student understanding?

The main themes that emerged from the interviews can be subsumed in two broad categories: aspects pupils identified that enable understanding (and hence suggest the respective extent of comprehensibility), and aspects that pupils mentioned that impede understanding:

Aspects that enable understanding	**Aspects that impede understanding**
Good, clear, precise, loud and intelligible pronunciation	Unclear pronunciation, "swallowing words"
Appropriate loudness / loud voice	Low voice / low loudness
Appropriate complexity of content	Highly complex content
Low to moderate articulation rate	High articulation rate
High fluency, as long as the pronunciation is clear, precise, accurate and intelligible	Disfluency or low fluency, frequent hesitation markers (pauses, fillers), repair (repetition, rephrasing, false starts), taking time over finding appropriate words, which can act as distractors
	High fluency if the subject matter is highly complex
High-frequency vocabulary	Low-frequency vocabulary
Articulating a message concisely	
Redundancy / repetition	Redundancy / repetition

Table 36 : Aspects relevant for enabling or impeding understanding

Before reporting on the results in detail, it is important to note that the literature distinguishes between intelligibility and comprehensibility. Munro and Derwing (1999) outline the difference as follows: while intelligibility is defined as a listener's *actual understanding* of L2 speech and hence denotes the ease or difficulty with which a listener understands such productions, comprehensibility denotes a listener's *perception of understanding*. Comprehensibility, then, is generally measured by a listener's rating of how easily they understand a language production (Munro & Derwing, 1999; Trofimovich & Isaacs, 2012). Since I aim to uncover discrete aspects of the field experts' perceptions and judgements of pre-service language teachers' L2 productions, the aspect under scrutiny is *comprehensibility*. Pronunciation as a main category

occupies a similarly prominent role when it comes to aspects the research participants value with reference to comprehensibility as when it comes to aspects they identify as markers of language proficiency. *Clarity, precision, accuracy* and *comprehensibility* emerged as sub-categories and all participants agreed in some form or another that these sub-categories played an important role in ensuring "good" pronunciation and hence comprehensibility. Indeed, they all considered "good" pronunciation to be essential for enabling understanding. Out of the sub-categories mentioned above, *clarity* seemed by far the most important factor for enabling understanding. *Clarity* does not seem to be grasped as an innate trait that a teacher either masters or not; instead, *clarity* can be enhanced through a lower articulation rate as B3 mentions below:

> *B3: Ehm, und si hets au schö so… gseit dass mo… VERSTOHT. Zum Bispil, da mit «shoulder», het si SCHÖ gseit, so richtig so langsam «SHOULDER», dass sis* [the pupil who is being addressed in the video-vignette] *verstoht. (Interview transcript B3, line 99-101)*

The preceding quote also serves as an illustrative example of the field experts' frequent references to "beautiful" pronunciation, which in their view seemed a key criterion for ensuring understanding. Aside from the reference to comprehensibility, the participants mainly equated "beautiful pronunciation" to an acoustically pleasing sound of speech, which was at times also connected to a particular native-sounding variety of English. "Beauty" was not in all instances exclusively referred to as a contributor to a pleasing acoustic experience; some participants also linked a "beautiful accent" to high comprehensibility (B4). Within the category of auditory perception, loudness (*sone*) received a lot more attention than pitch (*mel*). While pitch was barely mentioned, an appropriate volume was crucial for enabling understanding to four of five research participants. Quiet and timid-sounding speech particularly impedes understanding and implies insecurities, which some students mentioned would lead them to lose interest and get distracted. Hence, loudness was not only connected to self-confidence and to enabling understanding, but also to the teachers' ability of managing the classroom appropriately. In this context, some participants also explicitly stated that they preferred loud over quiet speech.

Worth mentioning here is also that redundancy and repetition were perceived both as enabling *and* as impeding understanding. Very much like the mediation strategies involved with *simplifying a text*, *amplifying a dense text* and *streamlining a text* outlined in the CEFR-CV (Council of Europe, 2018, 2020), redundancy, repetition and articulating a message concisely ultimately serve the same purpose: to "clarify meaning and facilitate understanding" (North & Piccardo, 2016, p. 31). Judging from the students' responses and from the descriptors of the

CEFR-CV (Council of Europe, 2018, 2020), employing the individual strategies is highly situational, contextual and content-dependent. If not applied suitably, they may fail to enable understanding. While one student clearly preferred if "complicated information" [*sic*] was streamlined instead of amplified and liked a concise presentation of the most important facts, others considered that repetition can enhance understanding – especially for weaker learners:

> *B4: Auch vor allem weil sie vielmal eigentlich das gleiche Thema lang fokussiert hat dann ist's einfacher für die Person [the weak learner]. (Interview transcript B4, line 246-247)*

Enabling understanding can be interpreted as one of the main goals of classroom communication. *Addressee-specificity* in particular is a category that deserves special mention, not only when it comes to assessing L2 teacher language competence, but also particularly when it comes to understanding the construct itself. One can go as far as placing it at the core of successful classroom communication and as one of the most central components of facilitating communication, understanding and learning (see chapter 2.3). If a teacher does not succeed in adapting their language to their students' abilities and needs, understanding and hence learning are unlikely to occur. By conducting semi-structured interviews and removing visual contextual information from the speech productions, students' responses were exclusively based on auditory information. Hence, these judgements provided particularly meaningful insights into whether a pre-service teacher succeeded in being comprehensible to the students. These insights, in turn, are valuable for gaining a more precise understanding of the construct *addressee-specificity* – a construct that has proven to be highly complex and multifaceted in nature (see chapters 2.5.4.3, 5.1.1, 5.2.1 and 5.2.2). The aspects mentioned below can lead to cautious interpretations that may aid further refinements to the construct itself.

Generally, all students displayed an awareness that teachers need to adapt their language to the students' proficiency and that the use of simple language in the classroom does not imply low language proficiency on the teachers' part. Moreover, the participants showed a particular awareness of a number of general strategies teachers can employ to facilitate understanding in the classroom, as summarised in the following table:

Strategies to facilitate understanding
Using high-frequency vocabulary ("basic" English / BICS)
Clear and intelligible pronunciation
Adequate, rather slow(er) pace
Amplifying a text (paraphrasing, explaining something in detail, or using German)

Streamlining a text (being concise)
Using additional (mediation) tools to visualise, e.g., blackboard
Being well prepared[34]

Table 37 : Strategies that facilitate understanding (Council of Europe, 2018, 2020)

As indicated in Table 36 and Table 37, vocabulary use seems to play a central role in teacher-student communication. All students indicated an awareness that there is a difference between "basic" or "simple", and "complex", "advanced", "difficult" or "challenging" vocabulary. The participants were also aware that teachers know both types and that they always have a choice between using simple or complex vocabulary to explain a subject matter:

> *B1: das ist eigentlich das BASIC halt. Es ist mehr gebildet, so. Das gebildete Englisch, so. Es gibt das basic Englisch und das gebildete Englisch und ich kann nur zum Beispiel das Basic und dazwischen sind zwei. (Interview transcript B1, line 138-140)*

Additionally, students agreed that using simple vocabulary substantially enables understanding:

> *B4: Dass sie eigentlich ein Deu/ eh ein Englisch geredet hat das ich eigentlich auch gu/ also gut verstehe nicht irgendwie solche englische [sic] Sachen die ich nicht verstehe so sch/ komplizierte Wörter. (Interview transcript B4, line 62-64)*

While B5 indicated a liking for being challenged by low-frequency vocabulary and not minding the increased difficulty, B3 manifested that the vocabulary needed to be adjusted depending on the complexity of the content. The more complex the content of an utterance, the more simple the vocabulary needs to be:

> *B3: Mmh, aso am Afang ischs recht guet gsi. Und döt halt mit de Fehler döt… ischs halt chli komplizierter worde. Döt chönnt me glaub chli eifacheri Wörter ssueche. (Interview transcript B3, line 162-163).*

This particular participants' view on vocabulary use indicated that this strategy seems just as situational and content-dependent as the mediation strategies mentioned above (cf. CEFR-CV, Council of Europe, 2018, 2020). From this point of view, the teacher does not only need to adapt their vocabulary use to their students' general language ability, but also adjust it to the

[34] To be «well prepared» seems to be a prerequisite for successful classroom practice and management rather than a strategy for facilitating understanding. Both concepts are closely linked, however, and were brought into close connection by the research participants.

relative complexity of the lesson's content. To explain a complex subject matter, therefore, the vocabulary needs to be simplified even more – potentially even so that the L1 comes into use:

> *B3: Nai, d Sproch isch recht guet gsi aso d Ussproch isch würkli guet gsi. Ehm, isch eifach dasses… für die isch da voll KLAR. Und mir müend da zersch mol so, LANGSAM, dass mers au verstönd und vilicht au so uf Dütsch go wel mr sus halt nöd verstönt. (Interview transcript B3, line 42-44)*

The high awareness of this participant supports the claim that the construct *addressee-specificity* is likely multifaceted, highly dynamic and complex. It highlights that there may be a broader range of constituents that make up the construct than assumed. Articulation rate and high fluency, for instance, are two constituents that could be linked to *addressee-specificity*. Much like is the case with vocabulary use, a teacher needs to differentiate between situations that require a lower speech rate and situations in which student understanding is not so prone to be impeded if the speech rate is high(er). Just as with vocabulary use, then, the articulation rate needs to be adjusted to the complexity of the subject matter: the more complex a topic, the more slowly a teacher needs to speak to ensure that students can follow. Overall, the students indicated that they preferred a moderate speech rate, i.e. not too fast, so that they can follow what is being said, but also not too slow, so that they get bored or distracted. A good middle ground was seen as contributing to increased clarity, and hence, to comprehensibility. While some participants connected speech rate to fluency, in the case of comprehensibility the few references made to fluency highlighted that adequately fluent and sufficiently loud speech contributed to enabling understanding. In line with the need for situational and content-dependent adjustment of vocabulary use and articulation rate, a complex subject matter may require an adjustment and implementation of a range of fluency-strategies. This highly context-specific and situational view may allow linking the fluency construct to mediation strategies such as *breaking down complicated information* (Council of Europe, 2018, 2020). For instance, two students mentioned that while high fluency is crucial for enabling understanding, it may impede it if complex content is communicated without being explained step-by-step and with regular pauses. The below examples refer to pre- and post-test takers' responses to test task 3, where they provided feedback to a fictional student on a piece of their writing:

Example 1:

B1 Wenn jemand flüssend redet bekommst du WORT FÜR WORT den ganzen Satz mit. Ja halt, bis sie fertig ist dann musst du das andere verarbeiten, mit… (Interview transcript B1, line 205-206)

Example 2:

B3 und denn het si de erscht Fehler erklärt. Und denn de zweit. Und denn het si nomol e langi Pause gmacht und DENN erscht de Dritt und nid alli anenand und denn ischs au nöd so verständlech. (Interview transcript B3, line 32-35)

Example 3:

I gits öppis Zuesätzlechs wo dänksch het si NID so guet gmacht.

B3: Aso iz einglech nöd usserd da halt das so nöd alles anenand isch sondern so, chli so Pause so (Interview transcript B3, line 60-63)

These points indicate the difficulty external assessors may experience when evaluating *addressee-specificity*. As mentioned above, they indicate that the criterion seems to be a highly context and content dependent, multifaceted construct that involves more than adapting one's language to the target group. Similar to how the CEFR-CV suggests that *adapting language* is merely one way of realising the mediation strategy to *explain a new concept*, the interview data support the idea that *addressee-specificity* may subsume a range of strategies. Indeed, it may need to be understood as a separate construct rather than an analytic evaluation criterion. One of these potential strategies can be linked to the mediation strategy *breaking down complicated information* (Council of Europe, 2018, 2020). Specifically, the research participants mentioned that it was important for the teachers to *break down complicated information* into manageable chunks, to take their time over explaining concepts, to double-check with the students whether they are following the discourse, and to do so slowly (see Table 36). Only then can pupils follow what is being said, especially when the subject matter is challenging. These aspects are mirrored in the CEFR-CV *scale breaking down complicated information*:

- breaking a process into a series of steps;
- presenting ideas or instructions as bullet points;
- presenting separately the main points in a chain of argument. (p. 127)

Apart from the above aspects regarding complexity with reference to classroom content and a need to adjust the choice of vocabulary, the speech articulation rate and the fluency strategies,

students did not mention grammatical or lexical complexity and accuracy. As previously stated, accuracy may be a category yet too demanding for lower secondary school students to assess. What they did recognise, however, were additional aspects that were not initially connected to facilitating understanding per se, but that may contribute to students' motivation and readiness to listen, learn, and interact with their teacher and the learning content. Even though these aspects may not seem crucial for ensuring understanding, they can be seen as factors related to a teachers' voice that may pave the way to understanding by contributing to a classroom environment that is conducive to learning.

Overall, and aside from the linguistic categories deduced from both the interview guide and the their own perceptions (see Appendix G), the research participants noticed and valued aspects such as self-confidence and a friendly-sounding voice, or displayed a sensitivity to teacher commitment. A major influence on how the research participants perceived an oral language production was the pre-service teacher's tone of voice. A friendly tone of voice was of particularly high value to three students. Two participants made a special mention that a tired, lethargic, demotivated and disinterested tone of voice had a demotivating effect on them. Thus, if a teachers' voice sounds lively and enthusiastic, it is more likely for these students to engage. Further, two students found it important and convincing if a teacher comes across as confident, has a confident tone of voice and leaves the impression that they have "a clear line". One student mentioned that the pre-service teacher should be as self-confident as a native speaker:

> *B1 Ehm… diese Person soll einfach… SICHERER sein mit sich selber. Dass sie… ehm… einfach… das redet als wär das… ihre Muttersprache. (Interview transcript B1, line 165-166)*

Finally, students also commented on the perceived dedication, commitment and positive reinforcement of a teacher. For example, they valued encouragement, praise and positive feedback and perceived it as motivating. Three participants stated that they find it agreeable and important when they see that a teacher is dedicated to their students and interested in their learning, and puts effort into supporting them:

> *B4 aso i/ sie hat so getönt als ob sie auch FÜR den Schüler wäre. Und auch, aso ja sie hat sich MÜHE gegeben, das hat man auch gemerkt. (Interview transcript B4, line 334-335)*

> *B1 was mir gefallen hat ist… ehm… es kam halt überzeugend vor, dass sie sich MÜHE gemacht hat. Das… hat mich so überzeugt aso man merkt dass… da Mühe drin steckt. (Interview transcript B4, line 55-57)*

Another way of recognising that the students and their success is in a teachers' interest was by noticing if a teacher was well prepared. The data analysis revealed, overall, that the research participants were very aware of and responsive to an addressee-specific expression. They seemed to recognise that a number of aspects seem to be involved in the construct *addressee-specificity* and that a range of different strategies can be employed to ensure understanding.

8.2.4. Findings RQ #3.3

I attempted to answer RQ #3.3. with the following aspects in mind: some participants made explicit reference to the cognitive demand the interview task placed on them, indicating that they found assessing teachers' language proficiency as challenging. One of the reasons they unanimously mentioned was that they considered their own proficiency as not advanced enough in order to complete the task reliably and successfully. In contrast, the other students found the entire task and the interview "easy". Finally, four out of five pupils loved the role reversal and had fun assessing teachers, even though it felt unusual to three of them. With these aspects in mind, I compared the students' judgements to the expert ratings:

> RQ #3.3: How do lower secondary school students' perceptions of pre-service English teachers' oral feedbacks in the target language English compare to those of trained experts in applied linguistics and English language teaching and learning?

The two language productions presented to the pupil research participants had received the following expert ratings:

Audio recording sample	Inhaltliche Ausführung der Aufgabe	Wortschatz: Wortwahl	Sprachliche Korrektheit	Aussprache & Betonung	Flüssigkeit	Kohäsion & Kohärenz	Adressaten-bezug: Lernende
9039	3	3	3	3	3	3	3
8842	3	2	2	1	1	1	2

Table 38 : Expert ratings of language production samples

As indicated in the previous section, not all PRLC-R criteria were referred to by the pupils: while *cohesion & coherence* and *task completion* had previously been excluded from the interview guide (see chapter 7.2) and hence were not discussed during the interview, *accuracy* (grammar) and lexis did not receive any attention. Instead, the pupils focused mainly on *pronunciation*, *fluency* and *addressee-specificity*. One way of collecting students' judgements of these criteria was by asking the pupils how strenuous they experienced it to understand the message, and how much effort they had to put in to access the information. Most participants reported that they had understood most of the content of both speech samples and could recapitulate some or most of the discrete content-related points of each recording. One participant avoided answering the question, which can either be interpreted as an indication that they could not be bothered or that they had not understood. Overall, both recordings seemed to be linguistically accessible to most participants. This finding is in line with the expert judgements on the criterion *addressee-specificity*, where both language production samples received a satisfactory to a very good expert rating. When elaborating in more detail, three pupils stated that they found the first recording 9039 easy to understand, and that the second recording 8842 was still comprehensible but required a lot of concentration. They reported that they understood most of the content, but that they found the second, quiet recording more strenuous to access. At first sight, these findings seem to indicate that there is somewhat a match between the pupils' and the expert raters' perceptions – which would stand in opposition with Gautschi's (2018) findings (see chapter 1.3).

To unpack the construct *addressee-specificity* in more detail, the research participants were also asked to assess how comprehensible they thought the respective speech productions would be for weak learners. Even though most pupils found both recordings more or less intelligible for themselves, their answers changed when they related them to weaker learners. Recording 9039, which received the highest expert rating in *addressee-specificity*, was estimated to be too complex to understand for weak learners, hence disagreeing with the expert rating. B1, B2, B3 and B5 agreed that this difficulty would be due to the frequent use of complex vocabulary and the relatively high complexity of content (grammar error correction). In contrast, participant B4 thought that this same recording would be relatively easily accessible to weaker learners – mainly due to the clear and precise pronunciation. Recording 8842 was assessed similarly: participants B2, B3, B4 and B5 thought that it would not be comprehensible for weak learners, primarily because of the unclear pronunciation and low quality of the overall recording. However, B4 added that the perceived redundancy within the language production could

enhance a weak learner's understanding. As the exception, B1 considered recording 8842 to be easy to understand, rendering it more comprehensible because of the lower articulation rate. In sum, the overall student judgement correlated with the expert rating slightly more strongly for recording 8842. A higher correlation between the pupils' and the expert judgments could be observed for *fluency*, where the pupils assessed the recordings almost congruently to the experts. *Pronunciation* as the final criterion revealed a few minor differences between pupil and expert ratings. While the field experts agreed with the expert judgement on recording 9039, the rating for *pronunciation* in recording 8842 was slightly higher. In sum, the pupils' perceptions of the teachers' general and profession-related language proficiency were near congruent. This suggests that even a weaker learner or a learner with significantly lower language skills may be able to assess a teachers' language competence. Where the pupils differed however was in their judgement of the extent to which a weak learner would understand the respective teacher. The data show that the pupils who score low on their own language proficiency and academic achievement assume that weaker learners would not understand either one of the teachers, whereas the stronger students differentiated more between the teachers and their ability to make themselves understood. Another interpretation could be that weaker students can assess this particular aspect more accurately because they can better relate based on their own experience as weak(er) learners. Either way, these insights reflect the heterogeneity of the classroom and the challenge L2 teachers face when communicating with their students in the target language. These preliminary insights are promising when it comes to providing an avenue for further research into teacher language competence and *addressee-specificity*.

8.2.5. Additional Insight: Feedback

Aside from language-related aspects, some research participants also mentioned criteria that were essential to ensuring effective feedback. They considered it important that feedback includes both positive and negative aspects. It seemed important to them to learn precisely what went wrong and what they can do to improve. For these participants, therefore, feedback needs to be constructively critical but also contain praise to motivate. One participant wanted feedback to be as concise and precise as possible. To this participant, feedback needs to be honest and start with a positive reference, which needs to be followed by a focus on one negative point at a time. Furthermore, students agreed that it is crucial that explanations are clear, precise, and have a golden thread. While it was stressed that the points made throughout a teacher's feedback

need to be supported with further elaboration, redundancy and repetition was considered negative and a distractor, as stated above (see chapter 8.2.3).

8.3. Reliability

After the presentation of the sub-study results, the present subchapter briefly elaborates on the reliability of the above-described coding process and the overall results. Reliability constitutes an important quality criterion in both quantitative and qualitative research (cf. Misoch, 2015). In the former, it relates to the "stability of findings across time, contexts, and research instruments" (O'Connor & Joffe, 2020, p. 4). In the latter, the idea of "a single, objective, external 'reality' the scientific method can directly reveal" (ibid. p. 4) is most often rejected. Instead, qualitative epistemologies recognise a given area of interest to be "composed of multiple perspectival realities that are intrinsically constituted by an individual's social context and personal history" (ibid. p. 4). Thus, qualitative and quantitative research address reliability from different perspectives. In both the main- and sub-study, *interrater reliability* (IRR) and *intercoder reliability* (ICR) are of central concern in relation to the reliability, validity and objectivity. While IRR and ICR have commonalities, they do differ in their definition and nature:

> ICR is a numerical measure of the agreement between different coders regarding how the same data should be coded. ICR is sometimes conflated with interrater reliability (IRR), and the two terms are often used interchangeably. However, technically IRR refers to cases where data are rated on some ordinal or interval scale (e.g., the intensity of an emotion), whereas ICR is appropriate when categorising data at a nominal level (e.g., the presence or absence of an emotion). Most qualitative analyses involve the latter analytic approach. (O'Connor & Joffe, 2020, p. 2)

Thus, while IRR considers the main-study, ICR is relevant for the qualitative sub-study. Chapters 2.5.4.4, 4.5.3, 5.1.1, 5.2.1 and 5.2.2 discuss the role of and results related to IRR in the main-study. The present section discusses the approach to and the role of *intercoder reliability* (ICR) in the sub-study. While the consideration of ICR is relatively common in qualitative research and most prevalent in content analysis, it is not necessarily ubiquitous (O'Connor & Joffe, 2020). Qualitative researchers disagree about the appropriateness of conducting ICR assessment and about the ways in which to administer such assessments (ibid. p. 3). Arguments in favour of reporting ICR encompass extrinsic and intrinsic concerns. The

former include aspects such as demonstrating the overall rigor of the research procedure, assessing the

> rigor and transparency of the coding frame and its application to the data, [or assuring that] the analysis transcends the imagination of a single individual. (ibid. p. 3)

Research studies that carry the potential for real-world consequences, increasing the robustness of the evidence-base by conducting an ICR assessment is valued. Intrinsic concerns of ICR, in contrast, encompass aspects such as motivating researchers to guarantee consistency when coding, promoting reflexivity and dialogue between collaborating researchers, or fostering discussions that contribute to the refinement of the coding frame (ibid.). In contrast, ICR can be interpreted as contradicting the interpretative agenda of qualitative research. Qualitative research does not aim to reveal universally objective facts. Instead, it places a high value on recognising, interpreting and communicating a diversity of perspectives (ibid.). By calculating ICR, this diversity and analytic necessity is compromised. Transparently documenting the analytic procedures, reinforcing findings with raw data, or providing substantial attention to and carefully describing deviant cases can, among others, strengthen the reliability argument of qualitative research (ibid.). Calculating ICR can, however, be interpreted to imply that "there is a single true meaning inherent in the data", and thus carry the risk to suggest false precision through a single numerical value (ibid. p. 5). The present sub-study is highly explorative in nature, small-scale and of low real-world repercussions. Its focus lies on exploring the diversity of perspectives lower-secondary school students offer with reference to their teachers' language proficiency rather than on translating their views into generalisable truths. To avoid inflating the insights of the present sub-study through a mathematical value and strongly emphasising the individuality of perspectives and interpretations, I did not double-code the interviews and thus did not to conduct an ICR assessment.

8.4. Limitations Sub-Study

A number of limitations need to be considered when consulting this sub-study. First, when conducting qualitative research, the objectivity of the researcher is crucial. Since I developed the interview guide, conducted the interviews and analysed the data myself and did not involve any additional, less biased coders, this requirement was difficult to meet and cannot be fully guaranteed. Second, potential Matthew effects cannot entirely be ruled out since the interviewees self-selected and volunteered to participate in the research. Hence, the sample was

not truly random. Third, as the interview guide was based on the PRLC-R, priming students by asking them about specific criteria could not be avoided – rather, it was a prerequisite in order to answer the research questions. This means that the interview questions related strongly to the area of interest rather than being completely open, and that therefore the data does not consist of subjective theories. While sensitising the research participants to the subject of interest is necessary to gain access to information that is related to the research questions, the issue remains of too strongly pre-empting, and hence too evidently influencing the research participants. In addition, it cannot fully be guaranteed that the students' judgements truly reflected their perceptions or whether they made random guesses or invented answers. Fourth, and in line with the previous point, cognitive overload may have significantly influenced the results. The task the research participants were asked to perform was highly complex, the interview questions were challenging, and the audio speech samples were relatively long (roughly two minutes each). Further, the speech samples were in a foreign language and the sound quality of the second sample was slightly compromised, all of which may have impeded the participants' ability to remember the content of the recordings, let alone to understand it. Finally, the research participants were asked to take on an unusual role. Instead of being the learner with little "power", they found themselves in reversed roles; they became the powerful evaluators. This role-reversal could have affected the validity and accuracy of their judgements. It remains unknown whether the research participants could judge the speech samples objectively and what internal reactions this role-reversal caused among the participants. In order to achieve a certain amount of perceived distance for the interviewees from their role as evaluators, I included a small number of questions following an advocatory approach (e.g., "To what extent would a weak learner understand this particular teacher?"). Judging from the self-reported enjoyment most students experienced in this role-reversal situation, however, this limitation may not have affected them negatively to the extent to skew the results significantly.

8.4.1. Ethical Considerations Sub-Study

As outlined in chapter 6.6, informed consent is essential for orderly procedures of qualitative and quantitative studies that involve human beings as research participants (Halse & Honey, 2005). Thus, full and accurate information about the research needs to be communicated to any research participant transparently. Based on the given information, informed consent means that autonomous subjects can make rational, informed choices with reference to their participation (ibid.). In this sub-study, the participants received an informative letter with

accurate, comprehensive details about the study, including both the inconveniences and the gains. The parents and students were asked to sign a consent form before the study started. Likewise, careful attention was devoted to concealing the identity of each of the students in the recording and analysis stages to the best of my ability. Additionally, the participants and their legal guardians were assured that any collected information (transcripts, audio files, etc.) can be removed from the data pool at any time and for no reason if so requested.

9

Discussion Sub-Study

The following chapter aims to discuss the results of the sub-study with reference to their implications in terms the construct of teacher language competence, and the potential value of the findings to L2 teacher education and L2 teaching practice. The sub-study was primarily concerned with examining the orientations and perceptions of "field experts" – in this case, lower secondary school students – and of what they value when asked to assess pre-service teachers' performance on independent and integrated test tasks. The goal thereby was to examine and identify appropriate criteria for the assessment of pre-service teachers' language proficiency according to indigenous performance criteria such as *addressee-specificity*. I consider pupils as "novice raters" and "field experts" when it comes to assessing teacher language competence and *addressee-specificity* in particular. I conducted semi-structured interviews and implicitly guided the pupil-judges via the interview questions as to the features of teacher language performance they should consider. The procedure resembles a small-scale rater cognition study and was designed to elicit the understandings of the target group with reference to constructs of oral teacher language competence.

9.1. General Discussion

Overall and when compiled, the pupils' judgements create almost a type of profile of aspects that, in their view, constitute high language proficiency, and aspects that enable understanding. The results suggest that not all factors the field experts associated with high language proficiency also facilitated their understanding. In other words, the synthesised judgements lend some support to the notion that high teacher language competence differs from high general language ability (Bleichenbacher et al., 2017; Bleichenbacher et al., 2014; Elder & Kim, 2014;

Freeman, 2017; Freeman et al., 2015; Freeman et al., 2009). The participants commented on language performances only. Nevertheless, their statements indicate that, depending on the context and the situation, L2 teachers may need to use a type of language that is not conventionally associated with high general language proficiency:

	High language proficiency	**Low language proficiency**	**Enabling understanding**	**Impeding understanding**
"Good" pronunciation (clear, precise, "comprehensible")	✓		✓	
"Bad" pronunciation (unclear)				✓
Appropriate loudness			✓	
Low voice / loudness				✓
High fluency	✓		✓	
Disfluency (hesitations, repetitions, fillers, false starts, pauses)		✓		✓
High-frequency words			✓	
Low-frequency words	✓			✓
Ability to articulate a message concisely	✓		✓	
Redundancy / repetition			✓	✓
High articulation rate	✓			✓
Low / moderate articulation rate			✓	
Error-free language production	✓			
Appropriate complexity of content			✓	
Highly complex content				✓

Table 39 : Summary of categories and their respective allocations

It is notable to emphasise the great variety of additional categories that emerged inductively from the transcripts during the category-building process. These inductive categories either supplemented and extended the existing categories, or diverged from them entirely. It is not uncommon for rater cognition studies to lead to insights that deviate from given initial criteria such as those presented here. Previous studies on rater cognition such as those conducted by Meiron (1998), Brown (2000) or Brown et al. (2005), where judges were asked to "naively" rate speech productions elicited by integrated and independent test tasks, reported that raters

did not solely focus on evaluation criteria that are commonly associated with assessing speaking (i.e. syntax and vocabulary). Rather, they tend to pinpoint aspects of performance that are not explicitly included in the scales, as well as aspects of communicative skills ranging from the use of communication strategies to discrete aspects of discourse such as its structure, organisation and content. Indeed, Brown et al. (2005) found that there are four major conceptual categories that raters commonly tend to when conducting their assessments, each being comprised of specific production features:

- phonology (encompassing pronunciation, intonation and rhythm),
- linguistic resources (mainly sophistication, complexity and accuracy),
- fluency (including hesitation and repair), and
- content.

Moreover, the emphasis placed on comprehensibility or clarity was particularly marked. These findings suggest that highly salient aspects of performance, therefore, not only included traditional linguistic resources (i.e. grammar and vocabulary), but also production features such as fluency and pronunciation. The findings of the present sub-study are in line with Meiron (1998), Brown's (2000) and Brown et al.,'s (2005) observations on rater cognition, notably when considering the striking predominance of pronunciation, fluency and additional aspects connected to communicative competence in the participants' statements. While field experts in both Meiron's (1998) and Brown's (2000) studies paid special attention test takers' fulfilment of functional demands of the test task (e.g., narration, description, etc.) and the test-takers' ability to cope with real-world (i.e. academic) demands (Brown et al., 2005), the participants in this sub-study made no such references. This can mostly be attributed to the fact that 1) participants were not explicitly guided to draw related conclusions and 2) that making such judgements may likely have exceeded their current L2 abilities. With *addressee-specificity* being a criterion of central importance when it comes to assessing L2 teacher language competence, a teachers' comprehensibility seems to inevitably be an indicator for high teacher language competence. Therefore, I will now specifically focus on comprehensibility in connection to the four major conceptual categories also identified by Brown et al. (2005) as a central construct of teacher language competence. The subsequent discussion aims to contribute to a deeper understanding of the research participants' perceptions with reference to teacher language competence and *addressee-specificity* in particular by taking a closer look at the findings with reference to *pronunciation*, *accuracy* and *fluency*. Even though the focus of this

sub-study lies on the participants' perceptions of speech and hence provides insights into the speech samples' comprehensibility, it is worth extending the term to intelligibility, its closely related sister-concept. The main reason for this broadening is to approach the constructs of teacher language competence and addressee-specificity with an extended perspective.

Pronunciation and variety / accent

Similar to Brown et al. (2005) who found that pronunciation was by far the largest and most predominantly mentioned subcategory of phonology when raters judged L2 speech performance, the research participants of the present sub-study also mentioned pronunciation most frequently in comparison to other language- and intelligibility-related aspects of performance. The prevalence of student references to pronunciation as one of the most central aspects that contributes to enabling understanding is in line with the traditional association of intelligibility with pronunciation (Trofimovich & Isaacs, 2012). Recent research, however, has revealed that intelligibility encompasses other linguistic criteria, such as discourse measures, lexical richness, or fluency measures (Trofimovich & Isaacs 2012). In other words, a range of linguistic features beyond pronunciation can affect a listener's comprehension of L2 speech (Isaacs, 2016). Moreover, accent and intelligibility have historically often been conflated in assessment scales for pronunciation (Isaacs, 2016; Isaacs & Harding, 2017), and the field has only recently started moving towards an intelligibility-approach (Levis, 2005, 2020). The intelligibility-approach suggests that L2 accent is a much narrower construct than previously assumed, where it was most strongly connected to the entire range of linguistic factors commonly referred to as pronunciation, such as word stress, rhythm and segmental production accuracy (Isaacs, 2016). This assumption is supported by the findings of Trofimovich and Isaacs' 2012 research study on disentangling accent from comprehensibility. The study examined the ratings of 60 laypeople and 3 experienced L2 teachers when evaluating 40 native French speakers' oral English performances against criteria related to accent and comprehensibility (i.e., phonology, fluency, lexis, grammar, discourse, etc.). The results show that accent and comprehensibility are indeed overlapping yet distinct constructs, and that both language variety and intelligibility as the broader conception of comprehensibility were associated with a broad range of speech measures (Trofimovich & Isaacs, 2012, p. 905). Accent was insofar disentangled from comprehensibility that it was exclusively associated with aspects of phonology (e.g., rhythm, segmental or syllable structure accuracy), whereas comprehensibility was predominantly related to grammatical accuracy and lexical richness. The

findings of the present sub-study relate to these results and the proposed extended approach to intelligibility insofar that the present field experts associated a broad range of linguistic features with overall comprehensibility (i.e. fluency, clear, loud and precise pronunciation, appropriate speech rate and the use of high-frequency vocabulary, see Table 39). Furthermore, the results support the more recent understanding of intelligibility where accent and pronunciation are disentangled from one another. Students' perceptions of variety or (L2) accent indeed align with the contemporary intelligibility-approach and the subdivision of the umbrella term pronunciation into (almost) its full range of constituents. While some participants tended to appreciate an agreeable-sounding and "pleasing" pronunciation, which was occasionally still connected to a native-like variety, they mostly did not consider such a variety as a prerequisite to ensuring intelligibility. In other words, they seemed to have an awareness that it is possible to be highly intelligible and highly proficient *and* to have a lingering L2 "accent". This finding is also in line with empirical evidence from studies showing that accent and intelligibility are partially independent dimensions (Trofimovich & Isaacs, 2012). For instance, speakers that are considered to have a strong L2 accent may still be fully intelligible, whereas unintelligible speakers are always rated as heavily accented (Derwing & Munro, 2009). Equally, the literature reports a tendency for weak relationships between accentedness ratings and intelligibility scores (at least of some raters) (Munro & Derwing, 1999).

Accuracy

Apart from pronunciation, empirical evidence (Trofimovich & Isaacs, 2012) commonly links comprehensibility to aspects of grammar and vocabulary in L2 speech. With reference to the impact of grammatical accuracy on L2 comprehensibility, previous research suggests that "listeners are distracted by grammatical errors from attending to the message in L2 speech, which makes comprehension more effortful" (Trofimovich & Isaacs, 2012, p. 913). This is manifested in findings that report negative effects of ungrammatical sentences on comprehensibility (Fayer et al., 1987; Varonis & Gass, 1982), connections between grammatical errors and reported comprehensibility (Munro & Derwing, 1999), and listeners' irritations caused by grammatical errors in L2 speech (Derwing et al., 2002). Despite these findings, it is interesting to note here that the students made almost no references to grammatical (and lexical) accuracy. Their post-interview comments indicated a high degree of difficulty when assessing these categories, which lends support to the assumption that their current

language proficiency was not yet sophisticated enough to be able to detect grammatical and lexical inaccuracies. This seems especially true if such inaccuracies are subtle:

Example 1:

I: Und zu erkennen, wie gut dass die im Englisch sind, wie war das für dich?

B1: Eben, das ist halt auch die andere Sache wo, ehm, nich, für mich SEHR schwierig ist, weil ich halt noch ein Schüler bin und noch nicht so hochgebildet bin wie SIE im Englisch. Ehm, da kann ich halt nicht sehr viel sagen dass das einfach für mich ist. So. Es braucht schon… die Zeit um das zu Verstehen mitzubekommen und dann kann ich erst die Rückgabe geben. Oder Rückmeldung, ja. (Interview transcript B1, line 301-306)

Example 2:

I: Was fandst du SCHWIERIG dran?

B2: Ehm, so bewerten halt. So, das fand ich ein bisschen schwierig (lacht). (Interview transcript B2, line 250-251)

Example 3:

I: Was het di EIFACH oder SCHWIRIG dünkt bim Beurteile? 00:32:49

B3: Ehm… öppis finde. Aso so zum Bispil ez bi de zweit Person öppis FINDE wo i dere Person cha ZEIGE wo si besser wür go. (Interview transcript B3, line 379-381)

Considering that the expert raters agreed that both pre-service teachers' lexical and grammatical accuracy were relatively advanced (sample 9039 received two 3s and sample 8842 two 2s in both categories respectively), the errors may indeed have been too subtle for lower secondary school students to notice or be deterred by. Nevertheless, with reference to lexical accuracy and richness of vocabulary, the pupils found the repeated use of high-frequency vocabulary more conducive to understanding. This contradicts Trovimovich and Isaac's 2012 findings that suggest that "richer, more varied lexical content of L2 speech (i.e. greater type frequency, or a larger number of unique content words) is associated with higher comprehensibility ratings" (p. 913). Instead, the pupils' comments complement and extend on previous research findings that

propose that the severity of rater judgements and the perceived quality of L2 speech can be impacted by L2 speakers' familiarity with L2 vocabulary (Munro et al., 1994). In addition, the students' statements add to the understanding that semantic context including lexis is an aspect that listeners rely on when evaluating L2 speech (Gass & Varonis, 1984). Despite this evidence, Trofimovich and Isaacs (2012) appeal to cautious interpretations of the vocabulary-comprehensibility link. Indeed, in their research they found strong associations between type frequency with fluency and discourse complexity measures. According to these findings, rich vocabulary "is also linked to more fluent word retrieval and articulation and to more complex discourse structure, and […] listeners may consider all these features in judging L2 comprehensibility" (ibid. p. 914). With these aspects in mind, and similar to the judgements of grammatical accuracy (or the lack thereof), the pupil's statements indicate that pre-service teachers were more likely to be understood when they employed more high-frequency vocabulary. These references to comprehensibility in connection with lexical richness seem to be connected to the pupils' own language proficiency. This is where the construct of *addressee-specificity* could be of relevance: it seems that if teachers are aware of the proficiency gap between themselves and the students and adjust the complexity of their language productions to their pupils' L2 proficiency, understanding (and therefore learning) will be eased. The following excerpts indicate such an awareness among the interviewees:

I: gab es etwas was dir besonders GUT gefallen hat?

B4: Dass sie eigentlich ein Deu/ eh ein Englisch geredet hat das ich eigentlich auch gu/ also gut verstehe nicht irgendwie solche englische Sachen [sic] die ich nicht verstehe so sch/ komplizierte Wörter.

I: Ja, also meinst du mit komplizierten Wörtern meinst du Wörter die du noch nicht kennst die ganz…

B4: Ja oder dass zum Beispiel ein Synonym, das viel schlimmer also nicht schlimm aber unverständlicher ist das meine ich kein. (Interview transcript B4, line 61-69)

In sum, comments on both grammatical and lexical accuracy were rare. While the participants mentioned grammatical accuracy only once and interpreted it as an indicator of high L2 proficiency, they did not make any obvious links between grammatical accuracy and

comprehensibility. With reference to lexical accuracy (and lexical richness), associations were more often made with comprehensibility than with proficiency.

Fluency

Students' perceptions of fluency and disfluency markers (hesitations and repair) also seem consistent with existing empirical evidence. For example, the research participants associated high fluency with the absence of hesitations, pauses, fillers and false starts, and disfluency with the presence of said markers as well as repair and occasionally slow articulation (see chapters 8.2.2 and 8.2.3). Indeed, Derwing et al. (2004) found similar results, showing that listener fluency judgments were associated with temporal measures (e.g., pausing, articulation rate, etc.). In their rater cognition study, Brown et al. (2005) found that fillers, hesitations and pauses were viewed negatively by the raters and raised concerns regarding the impact on intelligibility. In addition, repair fluency (repetitions, rephrasing, false starts) was occasionally evaluated positively because it was interpreted as evidence that test-takers could monitor their speech. However, an overall negative interpretation of repair fluency as a disruption to understanding prevailed among Brown et al.'s (2005) judges. Again, these findings correspond with the present research participants' judgements. Similar to the participants' references to and appreciation of teachers breaking down complex information to enable comprehension, the raters in Brown et al. (2005) also commented on test-takers' ability to break up speech productions into manageable chunks. Just like in the present study, this was regarded as natural and as facilitating understanding (Brown et al., 2005, p. 22).

9.2. Implications and Conclusions Sub-Study

Based on the above discussion of the research findings from the sub-study, the below sections present relevant implications with reference to consequences for further research in relation to defining, fostering and assessing teacher language competence, as well as consequences for the development of relevant instruments and tools to enrich both L2 teacher education and general L2 education.

9.2.1. Consequences for the Construct of Teacher Language Competence

The applied and action-oriented approach taken in this sub-study served to gain insight into the target group's perspectives on and judgements of L2 teacher competence as outlined in the PRLCP and PRLC-R. The findings provide first qualitative evidence that may inform further research into teacher language competence. For instance, the findings show that the current understanding of teacher language competence – just like the construct itself – is still on fuzzy ground. They also show that a teacher's ability to adapt her or his L2 language output to the respective L2 learners' current level of L2 proficiency, i.e. *addressee-specificity*, is pivotal to the construct of teacher language competence. This ability includes at least diagnostic competence on the teacher's part and her or his ability to react swiftly to the students' needs depending on the context, situation and lesson content by implementing particular strategies to facilitate understanding. Such strategies may include, among many others, to adapt one's L2 output depending on the complexity of the subject content by altering linguistic aspects such as pronunciation or articulation rate, making explicit and targeted use of paralinguistic features such as gestures and facial expressions, repeating and paraphrasing input, breaking down information into manageable chunks, using additional tools (e.g., blackboard), etc. Thus, the current understanding of *addressee-specificity* as a PRLC-R criterion may be enhanced considering the possibility of it being a distinct construct and investigating this by consulting the CEFR-CV mediation scales as potential complementary aspects. In addition, the further differentiation of teacher language competence could be achieved and the understanding the functioning *addressee-specificity* as an independent construct could be deepened when consulting stakeholders such as the target group in particular, and when conducting further research into the implementation of the CEFR-CV mediation scales in the L2 classroom. Thus, it seems promising to investigate a potential link or complementation between the PRCLP and PRLC-R and the CEFR-CV mediation scales. Furthermore, including students of the target level as field experts and thus as central stakeholders in the investigation process of implementing the PRLCP and PRLC-R in practice presents a holistic approach to the tools in practice. Particularly because the PRLCP and PRLC-R have essentially been developed to contribute to a more needs-oriented L2 teaching practice, it seems logical to investigate the view of the end-user, i.e. the addressee or "field expert". Ultimately, one may argue that only by asking the target group directly one can truly explore what essential aspects of teacher

language competence such as *addressee-specificity* entail and require. While the findings uncovered valuable insights, they are highly individual and qualitative in nature. Thus, more large-scale qualitative research needs to be conducted to investigate and potentially solidify the present findings. They are however highly promising when it comes to seeking a better understanding of the construct of teacher language competence and to aiming at disentangling its potential dimensions. When it comes to the validation of the PRLC-R as an assessment tool, insights from the target group as essential stakeholders are valuable and important. However, Gautschi (2018) speaks a word of caution with reference to interpreting the results of such studies:

> [W]hile the approach used is to take student assessments as evidence, it may be rightly asked to what degree the student perspective should be reflected in a rater tool. It is also recognised that student feedback has limitations especially in terms of the quality and reliability of responses (tickbox instruments may result in superficial, let's-get-this-over-with answers), or concerns regarding the comprehensiveness of information gathered. (Gautschi, 2018, p. 108)

These are legitimate concerns that highlight the importance of rigorous and carefully crafted research designs. The present findings are not representative but instead present a first approach to uncovering the multitude of perspectives of a multitude of stakeholders. Should such results be intended to provide sound validity evidence of an instrument such as the PRLC-R, more research of larger scale is necessary. Nevertheless, this sub-study shows that this path may indeed be worthwhile.

9.2.2. Consequences for the Research Instruments

I used the sub-study to answer the research questions by means of conducting semi-structured interviews and qualitative content analysis. These findings, albeit non-generalisable in nature, serve as verbal-report data with a high degree of authenticity generated by "field experts", who in this sub-study are the target group population. The results can thus assist potential future validation steps of the PRLC-R scales. With the PRLCP and the PRLC-R having been particularly designed to be applied in actual L2 teaching and learning settings, a direct link to authentic "real-world classroom contexts" is of particularly high relevance. However, as has been postulated by number of writers (Cumming et al., 2001, 2002; Fulcher, 1987; Matthews,

1990), rating scales used in the assessment of second-language proficiency often have no basis in actual performance. Rather, "[m]ost scales of language proficiency appear in fact to have been produced pragmatically by appeals to intuition, the local pedagogic culture and those scales to which the author had access" (Schneider & North, 1999). While the connection to authentic real-world contexts is often attempted to be established by including field experts in the scale development process, the bridge can often not be established reliably and evidently enough due to availability biases or matters of convenience. In conjunction with the lack of authenticity and the compromised ecological validity, Brindley (1991) draws attention to the crucial yet often unsatisfactorily answered question when consulting field experts for rating scale development: Who can be considered "expert"? Within the realm of L2 assessment, language teachers are the most commonly drawn upon. More recently, however, especially in LSP-testing, other possible stakeholders have been included. The OET, for example, famously involved a variety of health professionals (e.g., physiotherapy educators and supervisors) as domain experts when identifying indigenous language performance criteria to evaluate the clinical communication skills of trainee clinicians (Elder & McNamara, 2016). Other field experts might include test takers, people with whom learners will interact in the target context (i.e. lower-secondary school students) or "naïve" native speakers (Brown et al., 2005). By involving field experts with a unique understanding of the context of interest, the specific context of test use can therefore be conceptualised with a higher degree of authenticity, reliability and validity (Elder & McNamara, 2016). It is for these reasons that the findings presented in this chapter may prove to be of value to further differentiating and sharpening the still fuzzy concept of teacher language competence and the *addressee-specificity* dimension. Additionally, the results move the traditional focus of speaking assessment away from accuracy from an expert's perspective and strongly towards comprehensibility from the target group's perspective. With this shift in focus, the findings highlight the specificity and essence of LSP approach to speaking assessment and are promising in terms of adding to a deeper understanding of the constituents that make up teacher language competence. Since this qualitative sub-study is small-scale and only involved five research participants, the results provide an initial exploration for further research and need to be interpreted with caution. Additional larger-scale rater cognition studies involving lower secondary-school students as field experts are needed in order to review the present findings with reference to their reproducibility and usability in assessment scale development.

10 Conclusion

In this research study, I attempted to investigate the affordances and limitations of instruments that were developed to foster, measure and assess L2 teachers' profession-related language competences by the example of oral feedback skills. The overall aim was to investigate and explore the usability and application of these instruments when implemented in relevant real-world L2 teacher education and L2 classroom settings as well as in language testing contexts. By exploring the research questions, the study also aimed at gaining insights into the affordances and limitations of the tools to help refine the understanding of the underlying theoretical construct on a broader level, i.e. the concept of teacher language competence.

Overall, the action-oriented approach in both the main and sub-study to the implementation of the PRLCP and PRLC-R reveals an inherent fuzziness of the construct of teacher language competence as conceptualised in the PRLCP and the assessment criteria as captured in the PRLC-R. This overall finding aligns with the literature (see chapter 2.3) with reference to previous conceptualisations of the construct and at first sight is thus nothing entirely new. However, the PRLCP and PRLC-R are recent tools based on a much more refined and differentiated conceptualisation. Despite their much more concrete and applied nature, the results of the main-study indicate that it proved to be a challenge for L2 education experts to adequately understand and apply the researched criteria of the PRLC-R. Indeed, there was significant rater variation and problematic criteria discrimination within the PRLC-R assessment criteria. Especially the indigenous criterion *addressee-specificity* was particularly problematic. The thereof resulting low interrater reliability rendered the pre- and post-test data marginally usable for further analyses without the implementation of an MFRA to control the rater bias and variability. Interaction analyses showed that despite comprehensive rater trainings the raters did not apply the rating criteria in a uniform manner, that they were subject to gender bias, and that raters differed from one another significantly with reference to severity and leniency across the individual criteria and tasks. These findings point towards a stark need to revisit the existing criteria and further develop them. At the same time, the findings also show

that *addressee-specificity* seems to constitute a pivotal, if not the main feature of teacher language competence. Indeed, the criterion may in essence be a distinct construct given its seemingly multifaceted and multidimensional complex nature. It is thus not surprising that raters could not reach satisfactory agreement with reference to the PRLC-R criteria, given the inherent situational dependency, fuzziness, complexity and multidimensionality of the criteria. Controlling for these rater effects and implementing an MFRA to conduct the pre- and post-test analysis revealed that no significant treatment effects could be produced through the intervention study. Thus, it did not make a difference whether the participants practiced providing feedback by using the PRLC-R (E), whether they practiced providing feedback by devising and using their own assessment criteria (C1), or whether they did not practice providing feedback at all (C0). These results are likely to be attributed to a compromised functioning of the tools and rating processes; however, additional investigations are necessary to find clearer answers in this regard.

Further insights into the PRLC-R were gained in the sub-study. Indeed, the findings revealed promising avenues for further research into the PRLC-R criteria, especially with reference to the indigenous criterion *addressee-specificity*. At the same time, the sub-study contributed to highlighting the heterogeneity of the classroom and diversity of student perceptions on L2 teachers' language proficiency, i.e. their profession-related language competences in the language of instruction. Overall, the findings from both the main- and the sub-study support the notion that the construct of teacher language competence may contain more constituents than previously conceptualised, and that it is yet more complex than so far assumed. It is especially the criterion *addressee-specificity* that seems to be of central importance and influence within the broader construct. It also seems much more complex and multidimensional than conceptualised in the PRLC-R as there are indications of it to be comprised of a much bigger range of aspects, competences (i.e. diagnostic competence), strategies and linguistic features that reach beyond the audible and verbal. Indeed, *addressee-specificity* seems to be its own distinct construct that 1) may only be adequately assessed through a global scale and through the inclusion of field experts and a broad range of stakeholders, and that 2) needs a lot more research in order to be further differentiated.

While these results provide valuable insight into the (problematic) functioning of the PRLCP and PRLC-R as a framework of reference in Swiss L2 teacher education, it is important not to forget that successful, communicative and output-oriented language teaching practice likely depends on much more than high profession-related language competences of L2 teachers. The

many influencing facets of communicative language ability and thus teacher language competence are difficult to grasp let alone describe, and the ways in which they can be developed remain unclear and disputed (Caspari et al., 2016; Bachman & Palmer, 1996; Grum, 2012, McNamara, 1996). Thus, there is a limit to knowing how teacher language competence in particular can be fostered in teacher education, and to understanding how they influence L2 students' learning success (cf. Loder-Büchel, 2014). Aside from the inherent necessity to control for confounding variables in empirical (L2 teaching and learning) research to gain new, empirically sound knowledge on a subject matter, it is also important not to forget the overall meaning of the findings of applied research for all stakeholders involved. It is tempting to dismiss great efforts made in developing and providing materials for the real-world application, e.g., creating frameworks of reference such as the CEFR or the PRLCP, especially if they do not result in any statistically significant findings or immediate empirical validation. Even though both aspects are important and necessary to ensure evidence-based teacher education and to gain knowledge on treatment effects and the actual functioning of instruments in the wild, efforts such as the present research study are beneficial beyond statistical significance. The present research study presents an evidence-based approach to L2 teacher education interventions that has revealed insights to the underlying construct of the PRLCP – insights of which there need to be more in order to specify a construct before empirical studies building on or researching said construct can be conducted in the first place. That the construct of teacher language competence is a challenge to determine, let alone operationalise, has become clear already at the outset of this study. The many attempts of approximating a clear-cut definition of teacher language competence reflect this issue. The partial demarcation from general language competence marks teacher language competence as its own, very specific type of language competence. That teacher language competence reaches beyond language-related factors and amalgamates other large constructs such as diagnostic competence, PK, CK, or PCK, teacher language constitutes a very broad, multifaceted and rich construct that remains difficult to operationalise. The findings of the stark rater variability and differential rater functioning with reference to the PRLC-R underline this assumption. This study shows that there is still a lot of potential for further research and thus, that there is still a long way to go until a full, distinctive, reliable and valid definition of teacher language competence can be developed.

However, this study also shows that recent efforts such as the development of the PRLCP and PRLC-R, including the findings of this research, have contributed to identifying limitations of

the current conceptualisation of the construct and that thus constitute non-negligible steps towards slowly working advancing towards a more comprehensive understanding of it. Even if the tools are not yet empirically validated and at the stage they could be, applying them in the relevant context as tools to guide L2 teacher education interventions presents benefits, e.g., providing a framework of reference, contributing to evidence-based teaching and learning approaches in teacher education and the target level, harmonising l2 teacher education curricula across Switzerland, making assessment criteria more transparent or helping teacher educators with the conceptualisation of their modules. Similarly, this study has taken a step towards narrowing the gap between theory and practice within a scientific subject of inquiry where this gap presents a particularly pronounced field of tension. This has been achieved through adopting an approach of high ecological validity throughout both the main- and the sub-study with close proximity to the relevant context. In the same line of argumentation, this study also highlights the importance and benefits of including a broad range of perspectives in the development, trialling and refinement of teaching and learning materials – especially of those who are ultimately affected: the learners themselves. I conclude this dissertation by calling for more research at the intersection of theory and practice with reference to teacher language competence – research that holistically approaches a subject matter from a multitude of perspectives and strives for an evidence-based approach to teacher education – and by calling for conditions that allow output-oriented development projects to conduct accompanying empirical research within the project timeframe as discussed below.

10.1. Avenues for Further Research

The present study may be regarded as an initial exploratory study of the application of the PRLCP and PRLC-R in relevant L2 education and language testing contexts, and there is a range of suggestions for further research that result from the findings. The proposed avenues for further research are subsumed in three strands.

The first strand relates to research with reference to the PRLCP and the PRLC-R as tools to be implemented in L2 education to prepare aspiring L2 teachers to successfully teach, and to ultimately ensure that future L2 learners can be better supported throughout their language learning process. First, further investigations into the application of the PRLCP in teacher education and their effects on pre-service teachers' profession-related language competence development is needed. Such insights can contribute a) to further develop the profiles and b) to

devise teaching and learning materials to support pre- and in-service teachers' L2 teacher language competence development. Second, it is necessary to conduct additional in-depth investigations with reference to the PRLC-R and its applicability, usefulness and appropriateness for the assessment of profession-related language competences of L2 teachers. This may include research into further refining the existing criteria to increase their validity, reliability and usability. This may also include exploring whether more indigenous criteria can be derived; e.g. through collecting data on authentic workplace use of language, creating corpora and enriching the assessment criteria by more relevant markers of proficiency related specifically to the field (cf. Loeliger, 2015), or conducting field expert interviews with a range of stakeholders. Such an approach could contribute to empirically validating the assessment criteria, i.e. ensuring that the criteria are data-based instead of theory-based (Fulcher, 1987). Indeed, empirical data-based approaches are highly recommended. For example, conducting classroom observations and compiling corpora of authentic spoken and written language use in L2 classroom situations may be a valuable approach to enriching the PRLCP and PRLC-R, to determining their systemic relevance, and to creating targeted teaching materials for teacher education that are of close proximity to the target language use domain. Such a foundation could enable more data-driven language learning (DDL) in the context of L2 teacher education and L2 teacher practice. This recommendation is in line with Studer's (2019) call for implementing teaching methods related to DDL by means of corpora:

> DDL sollte im Wald der Sprachlernmethoden nicht (mehr) als exotisches Pflänzchen gelten, sondern als valable, ernsthaft zu prüfende Option des Sprachenlernens, mindestens des Lernens von sprachlichen Formen und Strukturen. (p. 20-21)

Such an approach could also contribute to enhancing the level of authenticity of assessment instruments, especially in the case of performance tests. This may even mean that eventually, the currently weak performance test could be further developed into a strong performance test with closer proximity to the specific purpose of teacher language competence (McNamara, 1996). Third, further research with reference to the PRLC-R may include investigations of the PRLC-R's potential use and affordances in formative assessment in L2 teacher education (e.g., learning with rubrics).

The second strand of avenues for further research constitutes investigations of the theoretical understanding of teacher language competence to further differentiate the overall multifaceted and vague construct. For example, much more research is needed into the dimensions that make up the construct and additional dimensions that may be included, such as for example diagnostic

competence. Furthermore, the (so far only) indigenous criterion *addressee-specificity* needs to be further researched. The fuzzy, vague and multifaceted construct needs to be clarified and subdivided into clear-cut categories. One approach could be to investigate the CEFR-CV (Council of Europe, 2018, 2020) mediation scales in combination with the PRLCP and PRLC-R. For example, the mediation scales could be transformed into assessment criteria and their application could be compared with the application of the PRLC-R to identify overlaps and complementing features. Another approach may constitute further investigations into rater perceptions and how they apply the PRLC-R. Qualitative research approaches could be employed to investigate such perceptions and subsequently possibly clarify the criteria.

The third and final strand concerns the research participants of any further studies related to the PRLCP and PRLC-R. The present research has only investigated the PRLCP and PRLC-R in combination with pre-service teachers. Another avenue for further research thus constitutes expanding the scope to including in-service teachers, teacher educators and policy makers – as the PRLCP and PRLC-R have been designed with the purpose of also being used in professional development.

References

(EDK), S. K. d. k. E. (2017). *Empfehlungen zum Fremdsprachenunterricht (Landessprachen und Englisch) in der obligatorischen Schule.* Retrieved from https://edudoc.ch/record/128697/files/empfehlungen_sprachenunterricht_d.pdf

Abdullah, M. Y., Hussin, S., & Ismail, K. (2019). Implementation of flipped classroom model and its effectiveness on English speaking performance. *International Journal of Emerging Technologies in Learning*, *14*(9), 130-147. https://doi.org/10.3991/ijet.v14i09.10348

ACTFL, A. C. o. t. T. o. F. L. (2012). *ACTFL proficiency guidelines 2012.* Retrieved 23 April 2019 from https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

Ajjawi, R., & Boud, D. (2017). Researching feedback dialogue: an interactional analysis approach. *Assessment & Evaluation in Higher Education*, *42*(2), 252-265. https://doi.org/10.1080/02602938.2015.1102863

Ajjawi, R., & Boud, D. (2018). Examining the nature and effects of feedback dialogue. *Assessment & Evaluation in Higher Education*, *43*(7), 1106-1119. https://doi.org/10.1080/02602938.2018.1434128

Ajjawi, R., & Regehr, G. (2019). When I say… feedback. *Medical Education*, *53*, 652-654. https://doi.org/https://doi.org/10.1111/medu.13746

Alderson, C. J., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129. https://doi.org/10.1093/a pplin/14.2.115

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). Macmillan.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation* (9th printing ed.). Cambridge University Press.

Allen, D., & Tanner, K. (2006). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE – Life Sciences Education*, *5*, 197–203.

Allwright, D. (2003). Exploratory practice: Rethinking practitioner research in language teaching. *Language Teaching Research*, *7*, 113–141.

ALTE. (2018). Guidelines for the development of Language for Specific Purposes tests. A supplement to the manual for language test development and examining. In. Cambridge: Association of Language Testers in Europe.

ALTE. (2020). ALTE principles of good practice. In. Cambridge: Association of Language Testers in Europe.

Andrews, S., & McNeill, A. (2005). Knowledge about language and the 'good language teacher'. In N. Bartels (Ed.), *Applied linguistics and language teacher education.* (pp. 159–178). Springer.

Arras, U. (2011). Mündliche Kompetenzen in der Fremdsprache fair messen. Überlegungen und Vorschläge zur Qualitätssicherung. *Babylonia*, *2*(11).

ASQ, A. f. S. Q. R. (2021). *QCAmap*. Retrieved 04.05.2021 from https://www.qcamap.org

Bacchus, R., Colvin, E., Bronwen Knight, E., & Ritter, L. (2020). When rubrics aren't enough: Exploring exemplars and student rubric co-construction. *Journal of Curriculum and Pedagogy*, *17*(1), 48-61. https://doi.org/10.1080/15505170.2019.1627617

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition*, *10*(2), 149-164. https://doi.org/10.1017/S0272263100007282

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2002a). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *21*, 5-18. https://doi.org/10.1111/j.1745-3992.2002.tb00095

Bachman, L. F. (2002b). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453-476. https://doi.org/10.1191/0265532202lt240oa

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34.

Bachman, L. F., & Dambök, B. (2018). *Language assessment for classroom teachers*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford University Press.

Bachmann, L., & Palmer, A. (2010). *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford University Press.

Bader-Lehmann, U. (2007). Teaching English in primary schools - Herausforderungen an die Sprachdidaktik und die Lehrerbildung. *Zeitschrift Beiträge zur Lehrerbildung*, *25*(2), 241-254.

Ball, D., Thames, M., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*, 389–407.

Beacco, J. C., Bouquet, S., & Porquier, R. (2004). *Niveau B2 pour le français (utilisateur / apprenant indépendant): Un Référentiel.* Didier.

Bearman, M., Boud, D., & Ajjawi, R. (2020). New directions for assessment in a digital world. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining university assessment in a digital world. The enabling power of assessment* (Vol. 7, pp. 7-18). Springer. https://doi.org/10.1007/978-3-030-41956-1_2

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper.* Educational Testing Service.

Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*, *8*, 215–241.

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014). Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, *43*(422–451).

Birri, T., & Smit, R. (2013). Lernen mit Rubrics. Kompetenzen aufbauen und beurteilen. *Pädagogik*, *3*, 34-39.

Björkman, B. (2011). English as a lingua franca in higher education: Implications for EAP. *Ibérica*, *22*, 79.

Black, P., & William, D. (1998). Assessment in classroom learning. *Assessment in Education: Principles, Policy, & Practice*, *81*(1), 7-74.

Bleichenbacher, L., Hilbe, R., Klee, P., Kuster, W., & Roderer, T. (2017). *Beurteilung berufsspezifischer Sprachkompetenzen von Lehrpersonen, die Fremdsprachen unterrichten (Projektresultate)*. https://www.phsg.ch/sites/default/files/cms/Forschung/Institute/Institut-Fachdidaktik-Sprachen/201711%20Beurteilung%20BSSK%20Produktebericht.pdf

Bleichenbacher, L., Kuster, W., Egli Cuenat, M., Klee, P., Roderer, T., Benvegnen, R., Schweitzer, P., Stoks, G., Kappler, D., & Tramèr-Rudolphe, M.-H. (2014). *Vergleich ausgewählter internationaler Sprachdiplome mit den berufsspezifischen Sprachkompetenzprofilen: Modelle und Empfehlungen für die Verwendung internationaler Sprachdiplome in der Aus- und Weiterbildung von Fremdsprachenlehrpersonen.*

Bleichenbacher, L., Kuster, W., Egli Cuenat, M., Klee, P., Roderer, T., Benvegnen, R., Schweitzer, P., Stoks, G., Kappler, D., & Tramèr-Rudolphe, M.-H. (2014c). Berufsspezifische Sprachkompetenzprofile für Lehrpersonen für Fremdsprachen: Schlussbericht zu den Projektetappen 3 und 4: 2012-2014. https://www.phsg.ch/forschung/projekte/berufsspezifische-sprachkompetenzprofile-fuer-lehrpersonen-fuer-fremdsprachen

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*, 3–13.

Blöte, A. W. (1995). Students' self-concept in relation to perceived differential teacher treatment. *Learning & Instruction*, *5*(3), 221-236.

Bond, L., Smith, R., Baker, W. K., & Hattie, J. A. (2000). *Certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study.* National Board for Professional Teaching Standards.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167. https://doi.org/10.1080/713695728

Boud, D., & Falchikov, N. (2007). *Rethinking assessment for higher education: Learning for the longer term.* Routledge.

Bower, M., Cavanagh, M., Moloney, R., & Dao, M. (2011). Developing communication competence using an online video reflection system: pre-service teachers' experiences. *Asia-Pacific Journal of Teacher Education*, *39*(4), 311–326. https://doi.org/10.1080/1359866X.2011.614685

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). SEAMEO Regional Language Centre.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy and Practice*, *11*(1), 7–26.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* ASCD.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 1-26. https://doi.org/10.1080/00131911.2014.929565

Brown, A. (2000). *An investigation of the rating process in the IELTS Speaking Module.* (Vol. 3). IELTS Austalia.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks.* Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-05-05.pdf

Brown, D. H. (2004). *Language assessment principles and classroom practices.* Pearson Education.

Burch, C. B. (1997). Creating a two-tiered portfolio rubric. *The English Journal*, *86*(1), 55-58.

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2016). *Nonverbal communication.* Routledge

Burke, B. M. (2013). Looking into a crystal ball: Is requiring high-stakes language proficiency tests really going to improve world language education? *Modern Language Journal*, *97*(2), 531–534

Burke, B. M. (2015). Language proficiency testing for teachers. *The Encyclopedia of Applied Linguistics*.

Burke, D. (2009). Strategies for using feedback students bring to higher education. *Assessment & Evaluation in Higher Education*, *34*(1). https://doi.org/10.1080/02602930801895711

Byram, M. (Ed.). (2003). *Intercultural competence.* Language Policy Division, DG IV – Directorate of School, Out-of-School and Higher Education, Council of Europe. https://rm.coe.int/16806ad2dd.

Byram, M. (2004). INCA Intercultural Competence Assessment-Portfolio Interkultureller Kompetenz. *Internet: www. incaproject. org*.

Byrd, D. R. (1994). Peer editing: Common concerns and applications in the foreign language classroom. *Die Unterrichtspraxis/Teaching German*, *27*, 119–123.

Cabrera-Solano, P. (2020). The use of digital portfolios to enhance English as a foreign language speaking skills in higher education. *International Journal of Emerging Technologies in Learning*, *15*(24), 159-176. https://doi.org/10.3991/ijet.v15i24.15103

Campbell, P. B. (1996). How would I handle that? Using vignettes to promote good math and science education. [Brochure]. In *American Association for the Advancement of Science.* Washington, D.C.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication.* (pp. 2-28).

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Candelier, M., Camilleri-Grima, A., Castellotti, V., de Pietro, J.-F., Lörincz, I., Meissner, F.-J., Schröder-Sura, A., & Noguerol, A. (2007). *A travers les langues et les cultures: Cadre de référence pour les approches plurielles des langues et des cultures (CARAP).* Conseil de l'Europe: Centre Européen Pour les Langues Vivantes (CELV). https://carap.ecml.at/Portals/11/documents/C4pub2007F_20080228_FINAL.pdf

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in higher education*, *31*(2), 219-233. https://doi.org/10.1080/03075070600572132

Carless, D. (2015). *Excellence in university assessment.* Routledge.

Carless, D. (2020a). Double duty, shared responsibilities and feedback literacy. *Perspectives on Medical Education*, *9*, 199–200. https://doi.org/10.1007/s40037-020-00599-9

Carless, D. (2020b). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*. https://doi.org/10.1177/1469787420945845

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in higher education*, *36*(4), 395-407. https://doi.org/10.1080/03075071003642449

Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*. https://doi.org/10.1080/13562517.2020.1782372

Caspari, D., Grünewald, A., Hu, A., Küster, L., Nold, G., Vollmer, H. J., & Zydatiss, W. (2008). *Kompetenzorientierung, Bildungsstandards und fremdsprachliches Lernen – Herausforderungen an die Fremdsprachenforschung* https://www.dgff.de/assets/Uploads/Kompetenzpapier-DGFF.pdf

Caspari, D., Klippel, F., Legutke, M., & Schramm, K. (2016). *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch.* Narr Francke Attempto.

Castañeda, M., & Rodríguez--González, E. (2011). L2 speaking self-ability perceptions through multiple video speech drafts. *Hispania*, *94*(3), 483–501. https://doi.org/10.1353/hpn.2011.0066

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369-383. https://doi.org/https://doi.org/10.1191/0265532203lt264oa

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, *36*(4), 498–506.

Chambless, K. S. (2012). Teachers' oral proficiency in the target language: Research on its role in language teaching and learning. *Foreign Language Annals*, *45*(1), 141–162. https://doi.org/10.111/j.1944-9720.2012.01183.x

Chan, C. S. C. (2017). Investigating a research-informed teaching idea: The use of transcripts of authentic workplace talk in the teaching of spoken business English. *English for Specific Purposes*, *46*, 72-89. https://doi.org/10.1016/j.esp.2016.12.002

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research.* (pp. 32-70). Cambridge University Press.

Chapelle, C., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language.* Routledge.

Cheung-Blunden, V., & Khan, S. R. (2018). A modified peer rating system to recognise rating skill as a learning outcome. *Assessment & Evaluation in Higher Education*, *43*(1), 58–67. https://doi.org/https://doi.org/10.1080/02602938.2017.1280721

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chong, S. W. (2021). Reconsidering student feedback literacy from an ecological perspective. *Assessment & Evaluation in Higher Education*, *46*(1), 92-104. https://doi.org/10.1080/02602938.2020.1730765

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Sage Publications Ltd.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*, 37–46

Connor, U., & Asenavage, K. (1994). Peer Response Groups in ESL Writing Classes: How much Impact on Revision? *Journal of Second Language Writing*, *3*, 257–276.

Coste, D., & Cavalli, M. (2015). *Education, mobility, otherness: The mediation functions of schools.* Council of Europe DGII – Directorate General of Democracy, Language Policy Unit.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. www.coe.int/lang-cefr

Crotty, M. (1998). *The foundations of social research.* Sage.

Cullen, R. (1998). Teacher talk and the classroom context. *ELT Journal*, *52*(3), 179-187. https://doi.org/10.1093/elt/52.3.179

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision-making and development of a preliminary analytic framework.* ETS.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*, 67-96.

Darsow, A., & Felbrich, A. (2014). Besondere Forschungsansätze: Experiment und Quasi-Experiment. In J. Settinieri, S. Demirkaya, A. Feldmeier, N. Gültekin-Karakoç, & C. Riemer (Eds.), *Einführung in empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache.* (pp. 229–241). Schöningh.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* (Vol. 7). UCLES/Cambridge University Press.

Davis, L. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* University of Hawaii at Manoa]. Honolulu.

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

De Grez, L., Valcke, M., & Roozen, I. (2009). The impact of an innovative instructional intervention on the acquisition of oral presentation skills in higher education. *Computers & Education*, *53*, 112–120. https://doi.org/10.1016/j.compedu.2009.01.005

Derwing, T., & Munro, M. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476-490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., Rossiter, M. J., & Ehrensberger-Dow, M. (2002). "They speaked and wrote real good": Judgements of non-native and native grammar. *Language awareness*, *11*, 84–99.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 665–679.

Doff, S., & Klippel, F. (2007). *Englisch Didaktik. Praxishandbuch für die Sekundarstufe I und II.* Cornelsen Verlag Scriptor.

Douglas, D. (1997). Language for specific purpose testing. In C. Clapham & D. Corson (Eds.), *Language Testing and Assessment* (pp. 111-119). Dordrecht, The Netherlands: Kluver Academic Publishers.

Douglas, D. (2000). *Assessing language for specific pruposes*. Cambridge University Press.

Douglas, D. (2010). *Understanding language testing*. Hodder Education.

Dresing, T., & Pehl, T. (2013). *Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende* (5 ed.). Eigenverlag.

Dubiner, D. (2018). 'Write it down and then what?': Promoting preservice teachers' language awareness, metacognitive development and pedagogical skills through reflections on vocabulary acquisition and teaching. *Language awareness*, *27*(4), 277-294 https://doi.org/10.1080/09658416.2018.1521815

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, *35*(4), 501-527.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2011). Facetten der Genauigkeit: Zur Reliabilität der Beurteilung fremdsprachlicher Leistungen [Facets of accuracy: On the reliability of foreign-language performance assessments]. *Deutsch als Fremdsprache*, *48*, 195–204.

ECML. (2017). Towards a Common European Framework of Reference for Language Teachers: Frameworks, Standards and Instruments. In: European Centre for Modern Languages.

EDK. (2004). Sprachenunterricht in der obligatorischen Schule: Strategie der EDK und Arbeitsprogramm für die gesamtschweizerische Koordination. In. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.

EDK. (2014). Lehrpersonen mit Unterrichtsbefähigung für Fremdsprachen: Stand und Entwicklungstendenzen. In. Bern: Schweizerische Konferenz der kantonalen Erziehungdirektoren.

EDK. (2017). Empfehlungen zum Fremdsprachenunterricht (Landessprachen und Englisch) in der obligatorischen Schule. In. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.

Egli Cuenat, M. (2014). Kompetenzorientierung in der fremdsprachlichen Bildung von Lehrpersonen: Berufsspezifisches Curriculum C1* im Projekt «Passepartout». *Beiträge zur Lehrerinnen-und Lehrerbildung*, *32*(3), 414-428.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, *18*(2), 149-170.

Elder, C., & Kim, S. H. O. (2014). Assessing teachers' language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment* (1 ed.). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla138

Elder, C., & McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing*, *33*, 153-174. https://doi.org/10.1177/0265532215607398

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* (pp. 261–287). Lawrence Erlbaum Associates, Inc.

Europe, C. o. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Press Syndicate of the University of Cambridge. https://rm.coe.int/1680459f97

Europe, C. o. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment - Companion volume with new descriptors*.

Faez, F., & Karas, M. (2019). Language proficiency development of non-native English-speaking teachers (NNESTs) in an MA TESOL program: A case study. *The Electronic Journal for English as a Second Language*, *22*(4), 1-16.

Fayer, J., M., & Krasinski, E. (1987). Native and non-native judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326.

Fenzl, T., & Mayring, P. (2017). QCAmap: eine interaktive Webapplikation für Qualitative Inhaltsanalyse. *Zeitschrift für Soziologie der Erziehung und Sozialisation*, *37*(3), 333-340.

Ferris, D. R. (2003). Responding to writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 119–140). Cambridge University Press.

Finch, A. E., & Sampson, K. (2004). *READI Oral Proficiency Criteria*. http://www.finchpark.com/courses/assess/READI-Oral-Proficiency-Criteria.pdf

Finch, J. (1987). The vignette technique in survey research. *Sociology*, *21*, 105–114.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions.* (3 ed.). Wiley

Foerster, H. v., & Pörksen, B. (1998). *Wahrheit ist die Erfindung eines Lügners. Gespräche für Skeptiker.* Carl-Auer-Systeme.

Freeman, D. (2015). *Educating second language teachers.* Oxford University Press.

Freeman, D. (2017). The case for teachers' classroom English proficiency. *RELC Journal*, *48*(1), 31 – 52. https://doi.org/10.1177/0033688217691073

Freeman, D., Katz, A., Gomez, P. G., & Burns, A. (2015). English-for-teaching. Rethinking teacher proficiency in the classroom. *ELT Journal 69*(2), 129 – 139.

Freeman, D., Orzulak, M. M., & Morissey, G. (2009). Assessment in second language teacher education. In A. Burns & J. C. Richards (Eds.), *The Cambridge guide to second language teacher education.* (pp. 77-90). Cambridge University Press.

Froidevaux, A.-C. (2012). *Writing skills for foreign language teachers. A case study of professional foreign language competences of teachers at lower secondary level* OAIster. http://worldcat.org/oclc/907961728/viewonline

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287-291

Fulcher, G. (2003). *Testing second language speaking*. Pearson.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment. An advanced resource book*. Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees, . *Language Testing*, *28*(1), 5-29. https://doi.org/10.1177/0265532209359514

Garbati, J. F., & Mady, C. J. (2015). Oral skill development in second languages: A review in search of best practices. *Theory and Practice in Language Studies*, *5*(9), 1763-1770. https://doi.org/10.17507/tpls.0509.01

Gartmeier, M., Bauer, J., Fischer, M. R., Hoppe-Seyler, T., Karsten, G., Kiessling, C., Möller, G. E., Wiesbeck, A., & Prenzel, M. (2015). Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. *Instructional Science*, *43*, 443 – 462. https://doi.org/10.1007/s11251-014-9341-6

Gartmeier, M., Bauer, J., Fischer, M. R., Karsten, G., & Prenzel, M. (2011). Modellierung und Assessment professioneller Gesprächsführungskompetenz von Lehrpersonen im Lehrer-Elterngespräch [Modeling and assessment of teachers' professional competence for parent-teacher conversations]. In O. Zlatikin-Troitschanskaia (Ed.), *Stationen Empirischer Bildungsforschung. Traditionslinien und Perspektiven* (pp. 412–426). VS.

Gautschi, C. (2018). EMI lecture quality parameters: the student perspective. *Bulletin VALS-ASLA*, *107*, 97-112.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment.* Falmer Press.

Glaboniat, M., Müller, M., Rusch, P., & Wertenschlag, L. (2005). *Profile Deutsch*. Langenscheidt.

Gläser-Zikuda, M. (2015). E-portfolio in higher education. In M. Spector (Ed.), *Encyclopedia of Educational Technology* (pp. 275–277). Sage.

Gläser-Zikuda, M., Feder, L., & Hofmann, F. (2020). Portfolioarbeit in der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König, M. Rothland, & S. Blömeke (Eds.), *Handbuch Lehrerinnen- und Lehrerbildung* (pp. 706–712). Klinkhardt. https://doi.org/10.35468/hblb2020-085

Goldman, R., Pea, R., Barron, B., & Derry, S. J. (2007). *Video research in the learning sciences.* (R. Goldman, R. Pea, B. Barron, & S. J. Derry, Eds.). Lawrence Erlbaum.

Gómez Sará, M. M. (2016). The influence of peer assessment and the use of corpus for the development of speaking skills in in-service teachers. *HOW Journal*, *32*(1), 103-128. https://doi.org/10.19183/how.23.1.142

Graddol, D. (2006). *English next: Why global English may mean the end of "English as a foreign language"*. British Council.

Grotjahn, R. (1987). On the methodological basis of introspective methods. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research.* (pp. 54–81).

Grum, U. (2012). *Mündliche Sprachkompetenzen deutschsprachiger Lerner des Englischen: Entwicklung eines Kompetenzmodells zur Leistungsheterogenität* (Vol. 45). Peter Lang.

Grum, U., & Legutke, M. K. (2016). Sampling. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch.* (pp. 78-90). Narr Francke Attempto.

Habermas, J. (1970). Towards a theory of communicative competence. *Inquiry*, *13*, 360–375.

Halliday, M. A. K. (1978). *Language as a social semiotic. The social interpretation of language and meaning.* Edward Arnold.

Halliday, M. A. K. (1985). *An introduction to functional grammar* (1 ed.). Edward Arnold.

Halse, C., & Honey, A. (2005). Unravelling ethics: Illuminating the moral dilemmas of research ethics. *Signs: Journal of Women in Culture and Society*, *30*(4), 2142-2161. https://doi.org/10.1086/428419

Harsch, C. (2016). Testen. In D. Caspari, F. Klippel, M. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch.* (pp. 204-217). Narr.

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77-89. http://www.unc.edu/courses/2007fall/jomc/801/001/HayesAndKrippendorff.pdf

Healey, J. F. (2012). *The essentials of statistics: A tool for social research.* Cengage Learning.

Helfferich, C. (2005). *Die Qualität qualitativer Daten. Manuel für die Durchführung qualitativer Interviews.* (4 ed.). Lehrbuch.

Hewlett, N., & Beck, J. M. (2006). *An introduction to the science of phonetics.* Lawrence Erlbaum.

Higgins, R., Hartley, P., & Skelton, A. (2002). The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in higher education*, *27*(1), 53-64. https://doi.org/10.1080/03075070120099368

Hoekje, B. (2016). "Language," "communication," and the longing for the authentic in LSP testing. *Language Testing*, *33*, 289-299. https://doi.org/10.1177/0265532215607921

Holliday, A. (2005). *The struggle to teach English as an international language.* Oxford University Press.

Hoo, H.-T., Deneen, C., & Boud, D. (2021). Developing student feedback literacy through self and peer assessment interventions. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.1925871

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*(4), 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Hughes, A. (1993). *Testing for language teachers.*

Hughes, R. (2011). *Teaching and researching speaking* (Second ed.). Pearson Education Limited.

Hughes, R., & Huby, M. (2002). The application of vignettes in social and nursing research. *Journal of Advanced Nursing*, *37*, 382–386.

Hung, S. T. A., & Huang, H. T. D. (2015). Video blogging and English presentation performance: A pilot study. *Psychological Reports*, *117*(2), 614–630.

Hunkeler, R. (2010). Die Überprüfung der berufsspezifischen Fremdsprachenkompetenz an Schweizer PH: Kurzüberblick und zwei praktische Beispiele. *Babylonia*, *3*(10), 58-63.

Hunkeler, R., Kuster, W., Manno, G., & Klee, P. (2009). *Umgang mit internationalen Sprachdiplomen an den Pädagogischen Hochschulen der Schweiz. Bericht zuhanden der EDK und der COHEP.*

Hyland, K., & Hyland, F. (2006). *Feedback on second language students' writing. Contexts and issues.* Cambridge University Press

Hymes, D. (1972). On communicative competence. In J. P. J. Holmes (Ed.), *Sociolinguistics* (pp. 269-293). Penguin.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. University of Pennsylvania Press.

Inbar-Lourie, O. (2013). Language assessment literacy. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0605

Isaacs, T. (2016). Assessing Speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (Vol. 131-146). DeGruyter Mouton.

Isaacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching*, *50*(3), 347-366. https://doi.org/10.1017/S0261444817000118

Jacoby, S. (1998). *Science as performance: Socializing scientific discourse through conference talk rehearsals.* University of California]. Los Angeles.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 485–514). American Council on Education and Macmillan.

Jones, R. L. (1979). Performance testing of second language proficiency. In E. J. Brière & F. B. Hinofotis (Eds.), *Concepts in language testing: some recent studies* (pp. 50-57). TESOL.

Jones, R. L. (1985). Second language performance testing: An overview. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing.* (pp. 15-24). University of Ottawa Press.

Joseph, G. E., & Brennan, C. (2013). Framing quality: Annotate video-based portfolios of classroom practice by pre-service teachers. *Early Childhood Education Journal*, *41*, 423–430. https://doi.org/10.1007/s10643-013-0576-7

Kaiser, G., Busse, A., Hoth, J., König, J., & Blömeke, S. (2015). About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Math Education*, *2*, 369-387.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kateri, M. (2014). *Contingency table analysis: Methods and implementation using R.* Springer

Kelly, M., & Grenfell, M. (2005). *European Profile for Language Teacher Education. A Frame of Reference*. University of Southampton.

Kelly, M., Grenfell, M., Allan, R., Kriza, C., & McEvoy, W. (2004). *European profile for language teacher education: A frame of reference*. European Commission Brussels.

Kennedy, A. S., & Lees, A. T. (2016). Preparing undergraduate pre-service teachers through direct and video-based performance feedback and tiered supports in early head start. *Early Childhood Education Journal*(44), 369–379. https://doi.org/10.1007/s10643-015-0725-2

Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*, *37*(3), 333–360. https://doi.org/10.1177/0265532219898635

Khan, Ö., & Taş, T. (2020). *On the models of communicative competence.* GLOBETSonline: International Conference on Education, Technology and Science,

Kırkgöz, Y. (2011). A blended learning study on implementing video recorded speaking tasks in task-based classroom instruction. *TOJET: The Turkish Online Journal of Educational Technology*, *10*(4), 1-13.

Kissau, S., & Algozzine, B. (2017). Effective foreign language teaching: Broadening the concept of content knowledge. *Foreign Language Annals*, *50*(1), 114-134. https://doi.org/10.1111/flan.12250

Kitney, A., & Morgan, S. (2019). *Authenticity: Turning theory into tasks.* IATEFL 2019 Pre-Conference Event, Liverpool. https://tea.iatefl.org/

Klaassen, R. (2001). *The international university curriculum: challenges in English-medium engineering education.* Technische Universiteit Delft]. Delft.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Renorth, H.-E., & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards - Eine Expertise.* Bundesministerium für Bildung und Forschung (BMBF).

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Knoblich, G., & Öllinger, M. (2006). Die Methode des Lauten Denkens, The Method of Thinking Aloud. In J. Funke & P. A. Frensch (Eds.), *Handbuch der Allgemeinen Psychologie – Kognition* (pp. 691-696.).

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, 1-25. https://doi.org/10.1177/0265532221994052

Knoch, U., & Macqueen, S. (2020). *Assessing English for professional purposes*. Routledge.

Königs, F. G. (2010). Faktorenkomlexion. In H. Barkowski & H.-J. Krumm (Eds.), *Fachlexikon Deutsch als Fremd- und Zweitsprache*. Francke.

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. https://doi.org/doi.org/10.1016/j.jcm.2016.02.012

Köroğlu, Z. Ç., & Çakır, A. (2017). Implementation of flipped instruction in language classrooms: An alternative way to develop speaking skills of pre-service English language teachers. *International Journal of Education and Development using Information and Communication Technology (IJEDICT)*, *13*(2), 42-55.

Krashen, S. (1981). *Second Language Acquisition and Second Language Learning.* Pergamon.

Krause, U.-M. (2007). *Feedback und kooperatives Lernen*. Waxmann.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, *30*, 61—70.

Krippendorff, K. (2004a). *Content analysis, an introduction to its methodology* (2 ed.). Sage Publications.

Krippendorff, K. (2004b). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411-433.

Kroll, B. (1990). *Second language writing: Research insights for the classroom.* Cambridge University Press.

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (4 ed.). Beltz Juventa.

Kuckartz, U., Dresing, T., Rädiker, S., & Stefer, C. (2008). *Qualitative Evaluation: Der Einstieg in die Praxis (2. aktualisierte Auflage)*. VS Verlag für Sozialwissenschaften.

Kuster, W., Klee, P., Egli Cuenat, M., Roderer, T., Forster-Vosicki, B., Zappatore, D., Kappler, D., Stoks, G., & Lenz, P. (2014). *Berufsspezifisches Sprachkompetenzprofil für Fremdsprachenlehrpersonen der Primarstufe und der Sekundarstufe I.* https://www.phsg.ch/forschung/projekte/berufsspezifische-sprachkompetenzprofile-fuer-lehrpersonen-fuer-fremdsprachen

Kvale, S. (2007). *Qualitative research kit: Doing interviews.* SAGE Publications Ltd. https://doi.org/10.4135/9781849208963

Lado, R. (1961). Language testing: The construction and use of foreign language tests. A teacher's book.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. https://doi.org/10.2307/2529310

Lantolf, J., Thorne, S. L., & Poehner, M. (2015). Sociocultural theory and second language development. In B. van Patten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 207-226). Routledge.

Lao-Un, J., & Khampusaen, D. (2018). *Using electronic portfolio to promote English speaking ability of EFL undergraduate students* ICETE 2018: 20th International Conference on Education, Teaching and E-learning, Tokyo, Japan.

Laurier, M., & Baker, B. (2015). The certification of teachers' language competence in Quebec in French and English: Two different perspectives? *Language Assessment Quarterly*, *12*, 10-28. https://doi.org/10.1080/15434303.2014.979349

Lázár, I. (2007). *Developing and assesing intercultural communicative competence: a guide for language teachers and teachers educators*. Council of Europe.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in higher education*, *23*(2), 157-172.

Lea, M. R., & Street, B. V. (2006). The "academic literacies" model: Theory and applications. *Theory into Practice*, *45*(4), 368-377.

Leeman, J. (2007). Feedback in L2 learning: Responding to errors during practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from linguistics and psychology* (pp. 111-137). Cambridge University Press.

Legutke, M. K. (2012). Fort- und Weiterbildung von Lehrkräften für Deutsch als Fremdsprache. In H.-J. Krumm, C. Fandrych, B. Hufeisen, & C. Riemer (Eds.), *Handbuch Deutsch als Fremd- und Zweitsprache.* (pp. 1351-1357). Walter de Gruyter.

Lehrpersonen, F. f. S. v. (2021). *Dritte Befragung zur Nutzung der Berufsspezifischen Kompetenzprofile für Sprachenlehrpersonen.*

Leki, I. (1990). Coaching from the margins: Issues in written response. In B. Kroll (Ed.), *Second language writing* (pp. 57– 68). Cambridge University Press.

Levi, T. (2012). *The effect of dynamic assessment on the performance of students in oral proficiency tests in English as a foreign language.* [Unpublished doctoral dissertation]. Tel Aviv University.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 367–377. https://doi.org/10.2307/3588485

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*(3), 310–328. https://doi.org/10.1075/jslp.20050.lev

Lewkowicz, J. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, *17*(1), 43-64. https://doi.org/10.1191/026553200669746135

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. Beltz Verlagsgruppe.

Linacre, J. M. (1998). Thurstone thresholds and the Rasch model. *Rasch Measurement Transactions*, *12*(2), 634-635.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2003). Estimating 50% cumulative probability (Rasch-Thurstone) thresholds. *Rasch Measurement Transactions*, *16*(4), 901.

Lindahl, K., & Baecher, L. (2016). Teacher language awareness in supervisory feedback cycles. *ELT Journal*, *70*(1), 28-38. https://doi.org/10.1093/elt/ccv047

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *20*(8), 15-21. https://doi.org/10.2307/1176232

Liu, J., & Sadler, R. W. (2003). The effect and affect of peer review in electronic versus traditional modes on L2 writing. *Journal of English for Academic Purposes*, *2*, 193–227.

Lockhart, C., & Ng, P. (1993). How useful is peer response? *Perspectives*, *5*(1), 17-29.

Loder-Büchel, L. (2014). *Association between young learners' English language performance and teacher proficiency and experience with English.* Université de Fribourg]. Fribourg.

Loeliger, M. (2013). Welchen handlungsorientierten, beruflichen Wortschatz brauchen Primarlehrpersonen für den Unterricht in Deutsch als Fremd-und Zweitsprache. *Schlussbericht Projektphase I. PH Fribourg*.

Long, M. (2005). *Needs analysis in second language learning.* Cambridge University Press.

Lunz, M. E., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Education and the Health Professions*, *13*, 425-444. https://doi.org/10.1177/016327879001300405

Luoma, S. (2009). *Assessing speaking* (5th printing ed.). Cambridge University Press.

Magnan, S. S. (1988). Grammar and the ACTFL Oral Proficiency Interview: Discussion and data. *Modern Language Journal*, *72*, 266-276.

Manias, E., & McNamara, T. (2016). Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing*, *33*, 235-249. https://doi.org/10.1177/0265532215608411

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal*, *44*(2), 117-121.

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes.* University of California at Los Angeles.

Mendonca, C. O., & Johnson, K. E. (1994). Peer review negotiations: revision activities in ESL writing instruction. *TESOL Quarterly*, *28*(4), 745-769.

Merriam-Webster.com. (2021). Feedback. In *Merriam-Webster.com*. Retrieved 15.04.2021, from https://www.merriam-webster.com/dictionary/feedback

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, *23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*, 741–749 https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Mettler, M. (2011). Berufsspezifische Sprachkompetenzen für Lehrpersonen, die Fremdsprachen unterrichten: Praxisbeobachtungen Teilprojekt II. *PHZ Luzern*.

Min, H.-T. (2005). Training students to become successful peer reviewers. *System*, *33*, 293–308. https://doi.org/10.1016/j.system.2004.11.003

Min, H.-T. (2016). Effect of teacher modeling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing*, *31*(31), 43-57. https://doi.org/http://dx.doi.org/10.1016/j.jslw.2016.01.004

Misoch, S. (2015). *Qualitative interviews.* De Gruyter.

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, *45*(4), 527–540. https://doi.org/10.1080/02602938.2019.1667955

Moodle. (2021). *Moodle PHSG website*. https://moodle.phsg.ch/

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology.* (pp. 745-783). Lawrence Erlbaum.

Müller, A., & Ditton, H. (2014a). *Feedback und Rückmeldungen. Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder.* Waxmann.

Müller, A., & Ditton, H. (2014b). Feedback: Begriff, Formen und Funktionen. In A. Müller & H. Ditton (Eds.), *Feedback und Rückmeldungen. Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder.* (pp. 11-28). Waxmann.

Muñoz Julio, W., & Ramírez Contreras, O. (2018). Transactional communication strategies to influence pre-service teachers' speaking skill. *Gist Education and Learning Research Journal*, *16*, 33-55.

Munro, M., J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, *11*, 253–266.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285–310.

Murillo-Zamoranoa, L. R., & Montanero, M. (2018). Oral presentations in higher education: a comparison of the impact of peer and teacher feedback. *Assessment and Evaluation in Higher Education*, *43*(1), 138-150. https://doi.org/10.1080/02602938.2017.1303032

Mutch, A., Young, C., Davey, T., & Fitzgerald, L. (2018). A Journey towards sustainable feedback. *Assessment & Evaluation in Higher Education*, *43*, 248–259. https://doi.org/10.1080/02602938.2017.1332154

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Nakatsuhara, F., Inoue, C., & Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*. https://doi.org/10.1080/15434303.2020.1799222

Narciss, S. (2006). *Informatives tutorielles Feedback. Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse.* Waxmann.

Nassaji, H. (2015). Qualitative and descriptive research: Data type versus data analysis. *Language Teaching Research*, *19*(2), 129–132. https://doi.org/10.1177/1362168815572747

Nelson, G., & Carson, J. G. (1998). ESL students' perceptions of effectiveness in peer response groups. *Journal of Second Language Writing*, *7*, 113–131.

Newby, D., Allan, R., Fenner, A.-B., Jones, B., Komorowska, H., & Soghikyan, K. (2007). *Europäisches Portfolio für Lehrpersonen in Ausbildung (PEPELF/EPOSA/EPOSTL)*. European Centre for Modern Languages (ECML). https://epostl2.ecml.at/Resources/tabid/505/language/de-DE/Default.aspx

Nicol, D. J. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501-517. https://doi.org/10.1080/02602931003786559

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199-218. https://www.reap.ac.uk/reap/public/Papers/DN_SHE_Final.pdf

Norris, J. M. (2013). Some challenges in assessment for teacher licensure, program accreditation, and education reform. *Modern Language Journal*, *97*(2), 554–560.

North, B. (2000). *The development of a common framework scale of language proficiency.* Peter Lang.

North, B. (2014). *The CEFR in practice* (Vol. 4). Cambridge University Press.

North, B., Mateva, G., & Rossner, R. (2013). *European Profiling Grid (EPG)*. European Commission. https://egrid.epg-project.eu/

North, B., & Piccardo, E. (2016). Developing illustrative descriptors of aspects of mediation for the CEFR. In *Common European framework of reference for languages: Learning, teaching, assessment.* Council of Europe.

Ntuli, E., Keengwe, J., & Kyei-Blankson, L. (2009). Electronic portfolios in teacher education: A case study of early childhood teacher candidates. *Early Childhood Education Journal*, *37*(2), 121–126.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1-13. https://doi.org/10.1177/1609406919899220

Ostrowski, M., Mouzourou, C., Danner, N., & Zaghlawan, H. (2012). Improving teacher practices using microteaching: Planful video recording and constructive feedback. *Young Exceptional Children*, *16*(1), 16–29.

Oxford, R. L., Lee, D. C., Snow, M. A., & Scarcella, R. C. (1994). Integrating the language skills. *System*, *22*(2), 257-268. https://doi.org/10.1016/0346-251X(94)90061-2

Panadero, E., & Jönsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Paulson, F., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership*, *48*(5), 60–63

Paulus, T. M. (1999). The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing*, *8*, 265–289

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Blackwell.

Pilkinton-Pihko, D. (2013). *English-medium instruction: Seeking assessment criteria for spoken professional English.* University of Helsinki].

Pill, J. (2019). *Authenticity in assessment: What does it mean?* IATEFL 2019 Pre-Conference Event, Liverpool. https://tea.iatefl.org/

Plakans, L. (2013). Assessment of integrated skills. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (Vol. 1, pp. 204-212). Wiley-Blackwell.

Port, R. F. (2007). The graphical basis of phones and phonemes. In O.-S. Bohn & M. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 349-365). John Benjamins.

Price, M., Handley, K., & Millar, J. (2011). Feedback: focusing attention on engagement. *Studies in higher education*, *36*(8), 879-896. https://doi.org/10.1080/03075079.2010.483513

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277–289. https://doi.org/10.1080/02602930903541007

Purpura, J. (2017). Assessing meaning. In E. Shohamy, L. G. Or, & S. May (Eds.), *Language testing and assessment, encyclopedia of language and education* (3 ed., pp. 33-61). Springer International Publishing AG.

Radford, B. W. (2014). *The effect of formative assessments on language performance* Brigham Young University].

Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second language research and teaching. *Canadian Modern Language Review*, *36*(2), 225–237. https://doi.org/10.3138/cmlr.36.2.225

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science*, *28*, 4-13. https://doi.org/10.1002/bs.3830280103

Ratner, C. (2002). *Cultural psychology: Theory and method*. Kluwer/Plenum.

Reimann, D. (2020a). *Methoden der Fremdsprachenforschung.* Narr Francke Attempto.

Reimann, D. (2020b). Sprachmittlung - Mediation im Fremdsprachenunterricht: Kompetenz und Bildungsziel. *Babylonia*, *3*(1), 10-21.

Richards, H., Conway, C., Roskvist, A., & Harvey, S. (2013). Foreign language teachers' language proficiency and their language teaching practice. *Language Learning Journal*, *41*(2), 231-246. https://doi.org/10.1080/09571736.2012.707676

Rodriguez-Gonzalez, E., & Castañeda, M. E. (2018). The effects and perceptions of trained peer feedback in L2 speaking: impact on revision and speaking quality. *Innovation in Language Learning and Teaching*, *12*(2), 120-136.

Ruiz-Primo, M. A., & Brookhart, S. M. (2018). *Using feedback to improve learning.* Routledge.

Salaberry, R. (2000). Revising the revised format of the ACTFL oral proficiency interview. *Language Testing*, *17*(3), 289–310.

Salem Al-Yaseen, W. (2020). Impact of jigsaw cooperative learning technique on enhancing Kuwait English language student-teachers' speaking skills. *The New Educational Review*, *61*(3), 119-130. https://doi.org/10.15804/tner.2020.61.3.10

Sandmann, A. (2014). Lautes Denken – die Analyse von Denk-, Lern- und Problemlöseprozessen. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (pp. 179-188).

Schneider, G., & North, B. (1999). *"In anderen Sprachen kann ich..." Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*.

Schneider, G., North, B., & Koch, L. (2001). *Portfolio européen des langues. Version pour jeunes et adultes–Europäisches Sprachenportfolio. Version für Jugendliche und Erwachsene–Portfolio europeo delle lingue. Versione per giovani e adulti–European Language Portfolio. Version for Young People and Adults.* Berner Lehrmittel-und Medienverlag. www.sprachenportfolio.ch

Schreier, M. (2014). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *15*. https://www.qualitative-research.net/index.php/fqs/rt/printerFriendly/2043/3635

Scott, V. M. (1996). *Rethinking foreign language writing.* Heinle & Heinle.

Seong, Y. (2017). Assessing L2 academic speaking ability: The need for a scenario-based assessment approach. *Columbia University Working Papers in Applied Linguistics & TESOL*, *17*(2), 36-40.

Shohamy, E. (1994). The use of language tests for power and control. In J. Alatis (Ed.), *Georgetown University round table on language and linguistics* (pp. 57-72). Georgetown University Press.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*, 4-14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

Slater, S. J. (1980). Introduction to performance testing. In J. E. Spirer (Ed.), *Performance testing: Issues facing vocational education.* (pp. 3-17). National Center for Research in Vocational Education.

Smit, R., & Birri, T. (2014). Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation*, *43*(5-13).

Sokolova, N. (2012). Teacher language competence description: Towards a new framework of evaluation. *Quality of Higher Education*, *9*, 75-97. https://doi.org/10.7220/2345-0258.9.3

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press.

Stemler, S., & Tsai, J. (2008). Best practices in interrater reliability. Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods.* (pp. 29–49).

Stiggins, R. J. (1991). Assessment literacy. *Ph Delta Kappan*, *72*(7), 534-539.

Stiggins, R. J. (1997). *Student-centered classroom assessment.* Merrill Education.

Studer, P. (2015). Coping with English: Students' perceptions of their teachers' linguistic competence in undergraduate science teaching. *International Journal of Applied Linguistics*, *25*(2), 183-201.

Studer, P. (2018). English in the age of comprehensive internationalization: Defining competence guidelines for teachers in higher education. *Bulletin VALS-ASLA*, *107*, 27-47.

Studer, T. (2019). Erntezeit!? Streiflichter auf aktuelle Arbeitsfelder der Fremdsprachendidaktik. In E. Peyer, T. Studer, & I. Thonhauser (Eds.), *IDT 2017, Band 1: Hauptvorträge.* (pp. 20-34). Erich Schmidt Verlag.

Sutton, P. (2012). Conceptualizing feedback literacy: knowing, being, and acting. *Innovations in Education and Teaching International*, *49*(1), 31-40. https://doi.org/10.1080/14703297.2012.647781

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*(3), 275-302. https://doi.org/10.1177/026553220101800302

Swissuniversities. (2015). *Empfehlungen zur Nutzung der berufsspezifischen Sprachkompetenzprofile für Lehrpersonen der Primarstufe und Sekundarstufe I im Rahmen der Aus- und Weiterbildung.* Retrieved from https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Kammern/Kammer_PH/Empf/EmpfehlungenAGFS_de.pdf

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *High Educ*, *76*, 467–481. https://doi.org/10.1007/s10734-017-0220-3

Taylor, L. (2005). Washback and impact. *ELT Journal*, *59*(2), 154–155. https://doi.org/10.1093/eltj/cci030

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*(21–36).

Team, R. (2020). *RStudio: Integrated development for R.* In RStudio, PBC. http://www.rstudio.com

Thonhauser, I. (2019). Welche fachdidaktische Kompetenz brauchen Lehrende? Einige Antworten im Blick auf die Textarbeit im Fremdsprachenunterricht. In E. Peyer, T. Studer, & I. Thonhauser (Eds.), *IDT 2017, Band 1: Hauptvorträge.* (pp. 163-174). Erich Schmidt Verlag.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529-554.

Tillmann, H. G., & Mansell, P. (1980). *Phonetik. Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozess.* Klett-Cotta.

Tinsley, H., & Weiss, D. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*, 358–376. https://doi.org/10.1037/H0076640

Tinsley, H., & Weiss, D. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling.* (pp. 95–124).

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905-916.

Tseng, S., & Yeh, H. (2019). The impact of video and written feedback on student preferences of English speaking practice. *Language Learning & Technology*, *23*(2), 145-158. https://doi.org/10125/44687

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of second language learning*, *9*(2), 147-170. https://mite.edu.hku.hk/f/acadstaff/399/Do_Secondary_L2_Writers_Benefit_from_Peer_Comments.pdf

Urquhart, L. M., Rees, C. E., & Ker, J. S. (2014). Making sense of feedback experiences: A multi-school study of medical students' narratives. *Medical Education*, *48*(2), 189–203. https://doi.org/10.1111/medu.12304

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition*, *4*, 114–136.

Vicente, S. (2012). Sprachpraktische Ausbildung angehender Fremdsprachenlehrer – Forschungsstand und Perspektiven. In T. Tinnefeld, I.-A. Busch-Lauer, H. Giessen, M. Langner, & A. Schumann (Eds.), *Hochschulischer Fremdsprachenunterricht – Anforderungen, Ausrichtung, Spezifik.* (pp. 77-90). htw saar.

Villamil, O. S., & De Guerrero, M. (1996). Peer revision in the L2 classroom: Social-cognitive activities, mediating strategies, and aspects of social behavior. *Journal of Second Language Writing*, *5*(1), 51–75.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Vygotsky, L. S. (1986). Thought and language. In. Cambridge: MIT Press.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Watzlawick, P. (1976). *How real is real?: Confusion, disinformation, communication.* Random House.

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written response. *Assessment and Evaluation in Higher Education*, *31*(3), 379–394.

Weidle, R., & Wagner, A. C. (1982). Die Methode des Lauten Denkens. In G. L. Huber & H. Mandl (Eds.), *Verbale Daten – Eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung.* (pp. 81-103.).

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies.* (pp. 45-66). Hogrefe.

Weir, C. J. (2005). *Language testing and validation. An evidence-based approach.* Palgrave Macmillan.

Weir, C. J., Vidaković, I., & Galaczi, E. (2013). *Measured constructs: A history of Cambridge English language examinations 1913-2012.* Cambridge University Press.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford University Press.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.

Winstone, N., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach* (1 ed.). Routledge. https://doi.org/10.4324/9781351115940

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, *52*(1), 17-37.

Wipperfürth, M. (2009). Welche Kompetenzstandards brauchen professionelle Fremdsprachenlehrer und-lehrerinnen? *ForumSprache, Ausgabe 2/2009*.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe.

Wulf, H. (2001). *Communicative teacher talk – Vorschläge zu einer effektiven Lehrersprache.* Max Hueber.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Language Testing*, *24*(2), 251–286. https://doi.org/10.1177/0265532207076365

Xi, X., & Sawaki, Y. (2017). Methods of test validation. In E. Shohamy, I. Or , & S. May (Eds.), *Language Testing and Assessment* (3 ed., pp. 193-209). Springer. https://doi.org/10.1007/978-3-319-02261-1_14

Yeh, H.-C., Tseng, S.-S., & Chen, Y.-S. (2019). Using online peer feedback through blogs to promote speaking performance. *Educational Technology & Society*, *22*(1), 1-14.

Young, J. W., Freeman, D., Hauck, M., Garcia Gomez, P., & Papageorgiou, S. (2014). *A design framework for the ELTeach program assessments.* (ETS Research Report Series, Issue. E. T. Service.

Zydatiss, W. (2005). *Bildungsstandards und Kompetenzentwicklung im Englischunterricht.*

Zydatiss, W. (2007). *Deutsch-Englische Züge in Berlin (DEZIBEL). Eine Evaluation des bilingualen Sachfachunterrichts an Gymnasien Kontext, Kompetenzen, Konsequenzen*. Peter Lang.

Appendices

The appendix contains a selection of the relevant material used and/or developed for the present study. In Appendix A, all PRLCP descriptors of Area of Activity 3 can be consulted. In Appendix B, two sample task specifications and corresponding test tasks are presented. Appendix C contains the complete PRLC-R including all assessment criteria and PLDs. The PRLC-R are followed by an excerpt of the rating manual and two annotated benchmarks of test task 5 in Appendix D. Appendix E contains the complete interview guide developed for the sub-study including all preparatory remarks for the interviewer. This section is followed by a sample interview transcript in Appendix F. The entire coding frame developed for the content analysis of the semi-structured, guided interviews of the sub-study including the coding rules conclude the appendices section in Appendix G.

A

PRLCP Area of Activity 3 Descriptors

Area of Activity 3: Assessing, Giving Feedback and Advising

Below are all descriptors of AoA 3 across all skills as conceptualised in the PRLCP:

3.1 *In the target language, the teacher is able to understand, analyse and assess written work by learners so as to be able to give differentiated feedback.*

a Analyse mistakes in a short text (e.g. brief report or short story) in order to point out typical sources of mistakes.

b Assess a written text (e.g. book review or entry for a website) according to certain criteria (e.g. content, range, linguistic means, precision of expression).

3.2 *In the target language, the teacher is able to understand, analyse and assess oral contributions by learners in order to be able to give differentiated feedback.*

a Pinpoint a learner's linguistic strengths and scope for improvement according to certain criteria (e.g. content, range, linguistic means, precision of expression, fluency and appropriateness of stylistic register) on the basis of a short monologue by the learner (e.g. short report with personal comments, text summary).

b Assess the oral contribution of two learners who are holding a dialogue (e.g. about a mobility exchange or about music).

3.3 *In the target language, the teacher is able to formulate tasks that enable him/her to assess a learner's linguistic abilities.*

a Formulate written questions to assess the comprehension of an audio or audiovisual document (e.g. song lyrics, film excerpt).

3.4 *In the target language, the teacher is able to give written instructions for self-assessment or assessment by fellow learners.*

a Give written guidelines for learners on how to assess work by a fellow learner (e.g. presentation or poster).

3.5 *In the target language, the teacher is able to give written feedback on work produced by learners.*

a Correct a written text by suggesting alternative formulations and other improvements.

b Give brief, personal and constructive feedback (praise, etc.) in written form to a written text (e.g. description of an experience, letter).

3.6 *In the target language, the teacher is able to give oral instructions for learners' self-assessment or assessment by fellow learners.*

a Encourage learners to assess their own progress and learning achievements.

b Teach learners how to use specific tools for assessing their own speaking skills (observation grid, European Language Portfolio, etc.).

3.7 *In the target language, the teacher is able to comment on the performance of a class.*

a Give overall feedback to the class at the end of a lesson (e.g. praise, encouragement, criticism, reprimand).

3.8 *In the target language, the teacher is able to conduct a dialogue that serves to assess a learner's ability.*

a Conduct a dialogue with a learner in order to assess, by certain criteria (e.g. content, range of vocabulary and grammatical means, precision of expression, fluency), his or her ability to participate in a conversation in the target language.

3.9 *In the target language, the teacher is able to give oral feedback on a learner's performance.*

a Give brief feedback on contributions by learners (e.g. praise, congratulations, criticism, indications on scope for improvement).

b Give constructive feedback on short learner presentations as encouragement for further studies.

3.10 *In the target language, the teacher is able to provide personalised information on the performance of individual learners.*

a Explain a learner's performance and progress clearly and succinctly to his or her parents who speak the target language.

3.11 *In the target language, the teacher is able to provide individual support to learners with explanations and advice to help them build up the ability of assessing themselves.*

a Support learners in their self-assessment by means of the European Language Portfolio.

3.12 *In the target language, the teacher is able to hold an advisory talk with learners with the aim of fostering their skills in a personalised manner.*

a Advise learners to set personal goals based on an assessment, and discuss these goals with them on an individualised basis.

b Address specific learning difficulties a learner is facing and discuss possible measures.

(Kuster et al., 2014, p. 13-14)

B Task Specifications and Test Tasks

Prüfungsaufgabe ‘Mündliche Produktion: 3.8’: Spezifikation Aufgabe 3

Angepasst nach Bachman und Damböck (2018).
Unterschiede zwischen realweltlicher Aufgabe und Testaufgabe jeweils hervorgehoben.

	Realweltliche Aufgabe	Testaufgaben
Kurzbeschreibung der Handlung, die in der «Wirklichkeit» bzw. in der Testaufgabe ausgeführt wird (Ausgewählte Deskriptor(en) und/oder Kurzbeschreibung der Handlung als Freitext)	Deskriptor 3.8: «Die Lehrperson kann in der Zielsprache ein Gespräch führen, das der Beurteilung der Kompetenzen der Lernenden dient». Die Lehrperson befindet sich in einem Einzelgespräch zum Thema «Reisen und kulturelle Unterschiede» zur formativen Beurteilung der mündlichen Sprachkompetenzen ihrer Schüler*innen. Die Lehrperson führt das Einzelgespräch und beurteilt anhand eines Beurteilungsrasters die mündlichen Sprachkompetenzen der Schüler*innen formativ. Die Lehrperson gibt dem/der Schüler/in eine mündliche Rückmeldung zu seinen/ihren mündlichen Sprachkompetenzen.	Deskriptor 3.8: «Die Lehrperson kann in der Zielsprache ein Gespräch führen, das der Beurteilung der Kompetenzen der Lernenden dient». Die Testperson erhält eine Beschreibung einer Klassensituation zur Kontextualisierung der Aufgabe. Sie schaut sich eine Videovignette an, in welcher sich eine Schülerin / ein Schüler zum Thema «angemessenes Verhalten in spezifischen kulturellen Kontexten» äussert. Sie muss der/m Lernenden in der Zielsprache eine kurze Rückmeldung zu seiner/ihrer Fähigkeit, sich in dieser Situation angemessen («Inhalt») und sprachlich korrekt («sprachliche Korrektheit») auszudrücken, geben. Diese Rückmeldung stützt sich auf ein Beurteilungsraster und erfolgt unmittelbar nach der Rezeption der Videovignette.

	Realweltliche Aufgabe	Testaufgaben
Konstrukt: Welche (Teil-)Kompetenzen werden für die Erfüllung der Aufgabe benötigt (sprachlich, fachdidaktisch, pädagogisch, vorausgesetztes thematisches/Weltwissen, IKT, …)? *[Nicht vergessen: Differenzen markieren]*	Inhaltlich: Wissen über Spracherwerb und kognitive Entwicklung, um einschätzen zu können, was die Lernenden sprachlich und inhaltlich vom Feedback verstehen Diagnostische Kompetenzen, um die mündlichen Sprachkompetenzen von Schüler*innen der Zielstufe Sek 1 einschätzen zu können. Diagnostische Kompetenzen, um das sprachliche Kompetenzniveau der Schüler*innen einschätzen und die eigenen sprachlichen Produktionen entsprechend gestalten zu können. Didaktisches und pädagogisches Wissen, um die Wirkung der formativen Beurteilung / des Feedbacks einschätzen und das Feedback entsprechend gestalten zu können. Sprachlich: Mündliche Kompetenzen, um adressatengerecht eine formative Beurteilung geben zu können (z.B. eigene Wortwahl, grammatische Komplexität, Tempo, Aussprache). Hohe sprachliche Korrektheit Klare Aussprache, hohe Flüssigkeit.	Inhaltlich: Genügend Wissen über mündliche Sprachkompetenzen der Zielstufe, um die beispielhafte Videovignette nachvollziehen und darauf reagieren zu können. Sprachlich: Mündliche Sprachproduktion in der Zielsprache, um der/dem Schüler/in basierend auf der beobachteten mündlichen Sprachproduktion und dem Beurteilungsraster eine Rückmeldung / formative Beurteilung zu den mündlichen Sprachkompetenzen (z.B. «Inhalt» und «sprachliche Korrektheit») geben zu können. Klare Aussprache, hohe Flüssigkeit. Hohe sprachliche Korrektheit. Sprachliches Repertoire, das es erlaubt, Wortwahl und grammatische Strukturen an die Kompetenzen der Lernenden anzupassen (z.B. eigene Wortwahl, grammatische Komplexität, Tempo, Aussprache).
Aufgabenmerkmale		

		Realweltliche Aufgabe	Testaufgaben
Setting	Raum, Material	Klassenzimmer, Lehrerzimmer, Nebenraum	Computerbasiert, schriftliche Kontextualisierung der Klassensituation, beispielhafte Videovignette zum 1x anschauen, Aufnahmemöglichkeit, ggf. mit Löschfunktion Material: Computer, Kopfhörer, Zugang zu Onlinetest
	Beteiligte (Personenkonstellation: wer kommuniziert mit wem?)	Lehrperson Schüler*innen der Sekundarstufe 1 (7.-9. Klasse)	Rollenspiel / Simulation: Testperson teilt einer beispielhaften, in der Videovignette dargestellten Klasse / einer/m beispielhaften, in der Videovignette dargestellten Schüler/in etwas mit. Die Antwort wird in ein Mikrofon eingesprochen und online gespeichert und ist monologisch.
	Benötigte Zeit	Vorbereitungszeit (studieren der Kriterien des Beurteilungsrasters), ca. 10-15 Minuten Einzelgespräch: Mündliche Produktion der/s Schülers/in beobachten – Notizen machen – formative Beurteilung basierend auf Kriterien eines Beurteilungsrasters geben Abhängig davon, wie lange das Einzelgespräch dauert, ca. 15-20 Minuten.	3-5 Minuten Vorbereitungszeit, um die Aufgabe zu studieren und die Videovignette zu schauen 0.5-2 Minuten Ausführzeit, um die eigene Sprachproduktion aufzunehmen
Input	Form des Inputs (z. B. Audio, Bilder, Fachtext, Lernertext, Items, z. B. Multiple-Choice mit 3 Optionen)	Beobachtbare Schülermeldung Absicht, einer/m Lernenden eine Rückmeldung zu ihren Sprachkompetenzen zu geben	Online Testaufgabe Kontextsetzung und Aufgabeninstruktionen Beispielhafte Videovignette Anweisung, einer/m Lernenden eine Rückmeldung zu ihren Sprachkompetenzen zu geben

		Realweltliche Aufgabe	Testaufgaben
	Merkmale der Sprache im Input (z. B. Komplexität, Wortschatz)	Beobachtbare Schülermeldung mit explizitem sprachlichen Input einer Schülerin / eines Schülers in der Zielsprache. Abhängig von der beobachtbaren Wortmeldung der Schülerin / des Schülers.	Beobachtbare Schülermeldung mit explizitem sprachlichen Input einer Schülerin / eines Schülers. Videovignette: sprachlicher Input in Form einer Schülermeldung in der Zielsprache.
	Länge (z. B. Wortanzahl, Dauer)	Abhängig davon, wie lange das Einzelgespräch dauert, ca. 3-5 Minuten.	Videovignette: ca. 20 Sekunden
	Erlaubte Themen	Einzelgespräch mit Schüler*innen der Zielstufe zur formativen Beurteilung ihrer mündlichen Sprachkompetenzen: Prinzipiell jegliche Art von Schülerantworten, welche in einem Klassenzimmer denkbar ist.	Einzelgespräch mit Schüler*innen der Zielstufe zur formativen Beurteilung ihrer mündlichen Sprachkompetenzen: Prinzipiell jegliche Art von Schülerantworten, welche in einem Klassenzimmer denkbar ist. Videovignette: kurze Schülerantwort auf eine vorgegebene, von der LP im Voraus gestellte Frage zu einem bestimmten Thema.
Erwarteter Output (Leistung, Antwort)	Form (z. B. mündliche Erklärung für Lernende, mündliche Erzählung, schriftliche Anweisung, korrigierte Fehler)	Mündliches Feedback / formative Beurteilung durch die Lehrperson in der Zielsprache, angepasst an die Schülermeldung und die Sprachkompetenz der Lernenden.	Aufnahme eines mündlichen Feedbacks durch die Testperson in der Zielsprache. Das mündliche Feedback geht auf die in der Videovignette dargestellte Schülerantwort ein und ist an die Sprachkompetenz der Lernenden angepasst.
	Länge (z. B. Wortanzahl, Dauer)	Ca. 3-5 Minuten. Abhängig von der Schülerantwort.	Ca. 30 Sek. – 2 Minuten. Abhängig von der Schülerantwort, keine Reaktion des Lernenden auf die Rückmeldung.

		Realweltliche Aufgabe	Testaufgaben
	Merkmale der Sprache (z. B. Komplexität, Wortschatz)	Adressatengerechte Äusserung in der Zielsprache, zusammenhängender Monolog zu unterschiedlichen Punkten. Rückmeldung zu mündlicher Produktion / Sprachkompetenz (z.B. «Inhalt» und «sprachliche Korrektheit») der/s Lernenden.	Adressatengerechte Äusserung in der Zielsprache, zusammenhängender Monolog zu unterschiedlichen Punkten. Rückmeldung zu mündlicher Produktion / Sprachkompetenz (z.B. «Inhalt» und «sprachliche Korrektheit») der/s Lernenden.
Erfassung der Leistung[35] (scoring)	Form der Bewertung (z. B. richtig/falsch, Punktzahl, Bewertung auf Skala)	Die Lernenden geben die Rückmeldung, dass sie (nicht) alles verstanden haben, z. B. durch Nicken, Nicht-Zurückfragen, Zurückfragen, Änderung des Arbeitsverhaltens etc.	Bewertung anhand spezifischer Qualitätsmerkmale zu «Sprechen monologisch» der Skalen des BSSK Beurteilungsrasters.
	Relevante bzw. bewertete Aspekte der Leistung	Relevanz der in der Rückmeldung angemerkten Punkte, z.B. zu verbesserungs-würdigen sprachlichen Aspekten (z.B. «Inhalt» und «sprachliche Korrektheit») mit Blick auf die dargelegte mündliche Sprachproduktion der Lernenden. Klarheit, Verständlichkeit aus Sicht der Lernenden (sprachlich und inhaltlich)	Aufgabenerfüllung Adressatengerechtigkeit Hohe Flüssigkeit und sprachliche Korrektheit Klare Aussprache
	Ermittlung des Ergebnisses		
	Genaues Vorgehen zur Ermittlung des Ergebnisses (Wer? Was? Wie?)		Die mündliche Produktion wird manuell von Experten anhand des BSSKP Beurteilungsrasters beurteilt.

[35] «Beurteilung» bei der realweltlichen Aufgabe kann z. B. bedeuten, dass eine Anweisung von SuS und FachdidaktikerInnen als klar empfunden wird.

		Realweltliche Aufgabe	Testaufgaben
	Konkrete Kriterien, ggf. Ratingskala		Inhaltliche Umsetzung der Aufgabe Wortschatz/Wortwahl Sprachliche Korrektheit Aussprache & Betonung Flüssigkeit: Tempo Kohäsion & Kohärenz Adressatenbezug: Lernende
	Zusammensetzung des Gesamtergebnisses bei der Aufgabe		Gewichtung der Sprache überwiegt bei den bewerteten Aspekten, solange ein Minimum an Inhalt vorhanden ist. Fokus auf sprachliche Produktion der Testperson in der Zielsprache, pädagogisches und didaktisches Wissen sowie Weltwissen wird nicht bewertet.

	Realweltliche Aufgabe	Testaufgaben
Fazit zur Konstruktvalidität	Zeigen Sie auf, inwiefern die Testaufgabe die Anforderungen der realweltlichen Aufgabe abbildet (Was bildet sie ab? Was verlangt sie zusätzlich? Was verlangt sie nicht?). Die Testaufgabe fokussiert auf das Sprachliche. Zwar verlangt die realweltliche Aufgabe das gleiche Mass an sprachlicher Beherrschung, kann aber inhaltlich weitergehen oder vom Arbeitsverhalten der Klasse / den Schülermeldungen anders ausfallen. Entsprechend muss die Lehrperson das Feedback angepasst auf die Situation formulieren und die relevanten Inhalte auswählen. Diese sind in der Testaufgabe vorgegeben. Die Lehrperson muss, anders als die Testperson, in der realweltlichen Aufgabe spontan auf die Klasse und mögliche Klassenreaktionen reagieren. Durch das computerbasierte Aufgabenformat kann im Test keine direkte Interaktion zwischen der Testperson und den Lernenden stattfinden, und so sind spontane Reaktionen der Testperson kaum abrufbar. Die Testperson kann sich beim Lösen der Aufgabe mehr Zeit nehmen und gegebenenfalls die Videovignette erneut anschauen. Zudem sind die Inhalte der mündlichen Produktion vorgegeben und sollten im Bereich des Verständlichen liegen. Der Testteilnehmende erhält kein unmittelbares Feedback durch die/den Lernenden. In der realen Welt könnte eine Lehrperson sehen, wenn etwas nicht gut verständlich ist. Die «Bewertung» findet in der realweltlichen Aufgabe durch den Lernenden statt, dessen Wahrnehmung und Bedürfnisse nur bedingt in der Testaufgabe wiedergegeben werden können. Der Fokus in der Testaufgabe liegt demnach stärker auf der sprachlichen Korrektheit als es in einer entsprechenden realweltlichen Situation der Fall wäre. Grund ist das Konstrukt: Der Lernende kann die Korrektheit der Lehrperson nicht (vollständig) beurteilen, sie ist für seinen Lernerfolg aber zumindest von Vorteil.	

Table 40 : Sample task specification test task 3

Testaufgabe «Mündliche Produktion» Sek 1: Frage 3

HF 3: Beurteilen, Rückmeldung geben und beraten: 3.8 ein Gespräch führen, das der Beurteilung der Kompetenzen der Lernenden dient.

Richtzeit Aufgabe: ca. 10 min / **Umfang Antwort:** 2-3 min Sprechzeit / **Zielgruppe:** 3. Klasse Oberstufe, Sekundarschule (erweiterte Anforderungen)

Situation:

Sie sie haben im Englischunterricht das Thema «Reisen und kulturelle Unterschiede» intensiv behandelt. Heute führen Sie mit Ihren Schüler*innen kurze Einzelgespräche zum behandelten Thema. Sie wollen die mündlichen Sprachkompetenzen Ihren Schüler*innen formativ beurteilen. Dabei konzentrieren Sie sich auf die Kategorien *Inhalt* und *sprachliche Korrektheit* auf dem Ihnen ausgeteilten Beurteilungsraster.

Im Einzelgespräch haben Sie Nathalie die Frage gestellt, in welchen Reise-Situationen man sich als Tourist besonders angemessen verhalten sollte.

Machen Sie sich mit dem Ihnen ausgeteilten Beurteilungsraster und der darunter stehenden Aufgabe vertraut. Schauen Sie sich danach das Video mit Nathalies Antwort an und machen Sie sich Notizen, auf welche Sie sich bei der anschliessenden Rückmeldung stützen.

Aufgabe:

Geben Sie Nathalie auf **Englisch** eine kurze Rückmeldung zu ihrer Fähigkeit, sich in dieser Situation **angemessen *(«Inhalt»)* und sprachlich korrekt *(«sprachliche Korrektheit»)*** auszudrücken.

Stützen Sie sich bei Ihrer Rückmeldung auf das Beurteilungsraster und Ihre Notizen.

Nehmen Sie Ihre Rückmeldung **auf Englisch** auf. Beachten Sie die angegebene Zeitvorgabe sowie das Niveau der Klasse. (**Sprechzeit** 2 - 3 min).

Figure 28 : Sample test task 3

Zusatzmaterial Frage 3 (E/F2.1 Formatives Beurteilungsgespräch «Reisen und kulturelle Unterschiede)

Kategorie	☹	😐	☺	Kommentare
Sprachliche Korrektheit	❑ Sie/er macht häufig Fehler.	❑ Sie/er macht manchmal Fehler.	❑ Sie/er macht nur sehr selten oder gar nie Fehler.	
Inhalt	❑ Die Wortmeldung ist inhaltlich unpassend	❑ Die Wortmeldung ist inhaltlich grundsätzlich passend.	❑ Die Wortmeldung ist inhaltlich treffend.	

Table 41 : Supplementing test material test task 3

Ihre Aufgabe für Ihre Rückmeldung an Nathalie:

Geben Sie Nathalie auf Englisch eine kurze Rückmeldung zu ihrer Fähigkeit, sich in dieser Situation angemessen («Inhalt») und sprachlich korrekt («sprachliche Korrektheit») auszudrücken.

Stützen Sie sich bei Ihrer Rückmeldung auf das Beurteilungsraster und Ihre Notizen.

Prüfungsaufgabe 'Mündliche Produktion: 3.9': Spezifikation Aufgabe 5

Angepasst nach Bachman und Damböck (2018).

Unterschiede zwischen realweltlicher Aufgabe und Testaufgabe jeweils hervorgehoben.

	Realweltliche Aufgabe	Testaufgaben
Kurzbeschreibung der Handlung, die in der «Wirklichkeit» bzw. in der Testaufgabe ausgeführt wird (Ausgewählte Deskriptor(en) und/oder Kurzbeschreibung der Handlung als Freitext)	Deskriptor 3.9: «Die Lehrperson kann in der Zielsprache mündliche Rückmeldungen zu Schülerleistungen geben». Die Lehrperson befindet sich in einer Klassensituation, in welcher Lernende Kurzvorträge zu einem vorgegebenen Thema halten. Die Lehrperson beobachtet den Kurzvortrag, macht sich auf einem Beurteilungsraster Notizen und erteilt eine mündliche Rückmeldung zum Beitrag. Die Lehrperson gibt den Lernenden eine mündliche Rückmeldung zu inhaltlichen und sprachlichen Aspekten des Kurzvortrags.	Deskriptor 3.9: «Die Lehrperson kann in der Zielsprache mündliche Rückmeldungen zu Schülerleistungen geben». Die Testperson erhält eine Beschreibung einer Klassensituation zur Kontextualisierung der Aufgabe. Sie schaut sich eine Videovignette an, in welcher Schüler*innen zum Thema «Britische Kultur» einen Kurzvortrag halten. Sie muss Lernenden in der Zielsprache eine kurze Rückmeldung auf ihre Beiträge (hier: Kurzvortrag) geben (z.B. loben, kritisieren, gratulieren, belohnen, auf Verbesserungsmöglichkeiten hinweisen usw.). Diese Rückmeldung bezieht sich auf einen konkreten inhaltlichen Aspekt des Kurzvortrags und zum Wortschatz und erfolgt unmittelbar nach der Rezeption der Videovignette.

	Realweltliche Aufgabe	Testaufgaben
Konstrukt: Welche (Teil-)Kompetenzen werden für die Erfüllung der Aufgabe benötigt (sprachlich, fachdidaktisch, pädagogisch, vorausgesetztes thematisches/Weltwissen, IKT, …)? *[Nicht vergessen: Differenzen markieren]*	Inhaltlich: Wissen über Spracherwerb und kognitive Entwicklung, um einschätzen zu können, was die Lernenden sprachlich und inhaltlich vom Feedback verstehen Diagnostische Kompetenzen, um die mündlichen Sprachkompetenzen und die Qualität eines Kurzvortrags von Schüler*innen der Zielstufe Sek 1 einschätzen zu können. Diagnostische Kompetenzen, um das sprachliche Kompetenzniveau der Schüler*innen einschätzen und die eigenen sprachlichen Produktionen entsprechend gestalten zu können. Didaktisches und pädagogisches Wissen, um die Wirkung der formativen Beurteilung / des Feedbacks einschätzen und das Feedback entsprechend gestalten zu können. Sprachlich: Mündliche Kompetenzen, um adressatengerecht eine formative Beurteilung geben zu können (z.B. eigene Wortwahl, grammatische Komplexität, Tempo, Aussprache). Hohe sprachliche Korrektheit Klare Aussprache, hohe Flüssigkeit.	Inhaltlich: Genügend Wissen über mündliche Sprachkompetenzen der Zielstufe, um die beispielhafte Videovignette nachvollziehen und darauf reagieren zu können. Sprachlich: Mündliche Sprachproduktion in der Zielsprache, um den Lernenden basierend auf dem beobachteten Schülerbeitrag und dem Beurteilungsraster eine Rückmeldung zu inhaltlichen und sprachlichen Aspekten (z.B. «Wortschatz», «grammatische Komplexität») geben zu können. Klare Aussprache, hohe Flüssigkeit. Hohe sprachliche Korrektheit. Sprachliches Repertoire, das es erlaubt, Wortwahl und grammatische Strukturen an die Kompetenzen der Lernenden anzupassen (z.B. eigene Wortwahl, grammatische Komplexität, Tempo, Aussprache).
Aufgabenmerkmale		

		Realweltliche Aufgabe	Testaufgaben
Setting	Raum, Material	Klassenzimmer	Computerbasiert, schriftliche Kontextualisierung der Klassensituation, beispielhafte Videovignette zum 1x anschauen, Aufnahmemöglichkeit, ggf. mit Löschfunktion Material: Computer, Kopfhörer, Zugang zu Onlinetest
	Beteiligte (Personenkonstellation: wer kommuniziert mit wem?)	Lehrperson Schüler*innen der Sekundarstufe 1 (7.-9. Klasse)	Rollenspiel / Simulation: Testperson teilt einer beispielhaften, in der Videovignette dargestellten Klasse / einer/m beispielhaften, in der Videovignette dargestellten Schüler/in etwas mit. Die Antwort wird in ein Mikrofon eingesprochen und online gespeichert und ist monologisch.
	Benötigte Zeit	Vorbereitungszeit (studieren der Kriterien des Beurteilungsrasters), ca. 10-15 Minuten Kurzvortrag beobachten – Notizen machen – Rückmeldung basierend auf Kriterien geben Abhängig davon, wie lange der Kurzvortrag dauert und wie viele Lernende vortragen, ca. 15-20 Minuten.	3-5 Minuten Vorbereitungszeit, um die Aufgabe zu studieren und die Videovignette zu schauen 0.5-2 Minuten Ausführzeit, um die eigene Sprachproduktion aufzunehmen
Input	Form des Inputs (z. B. Audio, Bilder, Fachtext, Lernertext, Items, z. B. Multiple-Choice mit 3 Optionen)	Beobachtbarer Kurzvortrag Absicht, Lernenden eine Rückmeldung zu ihrem Kurzvortrag zu geben	Online Testaufgabe Kontextsetzung und Aufgabeninstruktionen Beispielhafte Videovignette Anweisung, Lernenden eine Rückmeldung zu ihrem Kurzvortrag zu geben

		Realweltliche Aufgabe	Testaufgaben
	Merkmale der Sprache im Input (z. B. Komplexität, Wortschatz)	Beobachtbarer Kurzvortrag mit explizitem sprachlichen Input von Lernenden. Abhängig vom beobachtbaren Beitrag der Lernenden.	Beobachtbarer Kurzvortrag mit explizitem sprachlichen Input von Lernenden. Videovignette: sprachlicher Input in Form eines Kurzvortrags von Lernenden in der Zielsprache.
	Länge (z. B. Wortanzahl, Dauer)	Abhängig davon, wie lange der Kurzvortrag dauert und wie viele Lernende vortragen, ca. 3-5 Minuten.	Videovignette: ca. 1 Minute
	Erlaubte Themen	Kurze Rückmeldung auf Schülerbeiträge: Prinzipiell jegliche Art von Schülerbeiträgen, welche in einem Klassenzimmer denkbar sind.	Kurze Rückmeldung auf Schülerbeiträge: Prinzipiell jegliche Art von Schülerbeiträgen, welche in einem Klassenzimmer denkbar sind. Videovignette: Kurzvortrag zu einem von der LP im Voraus bestimmten Thema.
Erwarteter Output (Leistung, Antwort)	Form (z. B. mündliche Erklärung für Lernende, mündliche Erzählung, schriftliche Anweisung, korrigierte Fehler)	Mündliches Feedback durch die Lehrperson in der Zielsprache, angepasst an den Schülerbeitrag und die Sprachkompetenz der Lernenden.	Aufnahme eines mündlichen Feedbacks durch die Testperson in der Zielsprache. Das mündliche Feedback geht auf den in der Videovignette dargestellten Schülerbeitrag ein und ist an die Sprachkompetenz der Lernenden angepasst.
	Länge (z. B. Wortanzahl, Dauer)	Ca. 3-5 Minuten. Abhängig vom Schülerbeitrag.	Ca. 30 Sek. – 2 Minuten. Abhängig vom Schülerbeitrag, keine Reaktion des Lernenden auf die Rückmeldung.
	Merkmale der Sprache (z. B. Komplexität, Wortschatz)	Adressatengerechte Äusserung in der Zielsprache, zusammenhängender Monolog zu unterschiedlichen Punkten. Rückmeldung zum Schülerbeitrag.	Adressatengerechte Äusserung in der Zielsprache, zusammenhängender Monolog zu unterschiedlichen Punkten. Rückmeldung zu inhaltlichen und sprachlichen Aspekten des Schülerbeitrags (z.B. «Wortschatz» oder «grammatische Komplexität»).

		Realweltliche Aufgabe	Testaufgaben
Erfassung der Leistung[36] (scoring)	Form der Bewertung (z. B. richtig/falsch, Punktzahl, Bewertung auf Skala)	Die Lernenden geben die Rückmeldung, dass sie (nicht) alles verstanden haben, z. B. durch Nicken, Nicht-Zurückfragen, Zurückfragen, Änderung des Arbeitsverhaltens etc.	Bewertung anhand spezifischer Qualitätsmerkmale zu «Sprechen monologisch» der Skalen des BSSK Beurteilungsrasters.
	Relevante bzw. bewertete Aspekte der Leistung	Relevanz der in der Rückmeldung angemerkten Punkte, z.B. zu sprachlichen und inhaltlichen Aspekten (z.B. «Wortschatz» oder «grammatische Komplexität») mit Blick auf den dargelegten Kurzvortrag der Lernenden. Klarheit, Verständlichkeit aus Sicht der Lernenden (sprachlich und inhaltlich)	Aufgabenerfüllung Adressatengerechtigkeit Hohe Flüssigkeit und sprachliche Korrektheit Klare Aussprache
	Ermittlung des Ergebnisses		
	Genaues Vorgehen zur Ermittlung des Ergebnisses (Wer? Was? Wie?)		Die mündliche Produktion wird manuell von Experten anhand des BSSKP Beurteilungsrasters beurteilt.
	Konkrete Kriterien, ggf. Ratingskala		Inhaltliche Umsetzung der Aufgabe Wortschatz/Wortwahl Sprachliche Korrektheit Aussprache & Betonung Flüssigkeit: Tempo Kohäsion & Kohärenz Adressatenbezug: Lernende

[36] «Beurteilung» bei der realweltlichen Aufgabe kann z. B. bedeuten, dass eine Anweisung von SuS und FachdidaktikerInnen als klar empfunden wird.

		Realweltliche Aufgabe	Testaufgaben
	Zusammensetzung des Gesamtergebnisses bei der Aufgabe		Gewichtung der Sprache überwiegt bei den bewerteten Aspekten, solange ein Minimum an Inhalt vorhanden ist. Fokus auf sprachliche Produktion der Testperson in der Zielsprache, pädagogisches und didaktisches Wissen sowie Weltwissen wird nicht bewertet.
Fazit zur Konstruktvalidität	Zeigen Sie auf, inwiefern die Testaufgabe die Anforderungen der realweltlichen Aufgabe abbildet (Was bildet sie ab? Was verlangt sie zusätzlich? Was verlangt sie nicht?). Die Testaufgabe fokussiert auf das Sprachliche. Zwar verlangt die realweltliche Aufgabe das gleiche Mass an sprachlicher Beherrschung, kann aber inhaltlich weitergehen oder vom Arbeitsverhalten der Klasse / den Schülermeldungen anders ausfallen. Entsprechend muss die Lehrperson das Feedback angepasst auf die Situation formulieren und die relevanten Inhalte auswählen. Diese sind in der Testaufgabe vorgegeben. Die Lehrperson muss, anders als die Testperson, in der realweltlichen Aufgabe spontan auf die Klasse und mögliche Klassenreaktionen reagieren. Durch das computerbasierte Aufgabenformat kann im Test keine direkte Interaktion zwischen der Testperson und den Lernenden stattfinden, und so sind spontane Reaktionen der Testperson kaum abrufbar. Die Testperson kann sich beim Lösen der Aufgabe mehr Zeit nehmen und gegebenenfalls die Videovignette erneut anschauen. Zudem sind die Inhalte der mündlichen Produktion vorgegeben und sollten im Bereich des Verständlichen liegen. Der Testteilnehmende erhält kein unmittelbares Feedback durch die/den Lernenden. In der realen Welt könnte eine Lehrperson sehen, wenn etwas nicht gut verständlich ist. Die «Bewertung» findet in der realweltlichen Aufgabe durch den Lernenden statt, dessen Wahrnehmung und Bedürfnisse nur bedingt in der Testaufgabe wiedergegeben werden können. Der Fokus in der Testaufgabe liegt demnach stärker auf der sprachlichen Korrektheit als es in einer entsprechenden realweltlichen Situation der Fall wäre. Grund ist das Konstrukt: Der Lernende kann die Korrektheit der Lehrperson nicht (vollständig) beurteilen, sie ist für seinen Lernerfolg aber zumindest von Vorteil.		

Table 42 : Sample task specification test task 5

Testaufgabe «Mündliche Produktion» Sek 1: Frage 5

HF 3: Beurteilen, Rückmeldung geben und beraten: 3.9 mündliche Rückmeldungen zu Schülerleistungen geben.

Richtzeit Aufgabe: ca. 4 min / **Umfang Antwort:** 1-2 min Sprechzeit / **Zielgruppe:** 3. Klasse Oberstufe, Realschule (allgemeine Anforderungen)

Situation:

Im Englischunterricht behandeln Sie zur Zeit das Thema «Britische Kultur» mit speziellem Fokus auf den dazugehörigen Wortschatz. Für eine summative Prüfung haben Sie Ihre Schüler*innen beauftragt, in Zweiergruppen Kurzvorträge zu einem Teilaspekt des Themas zu halten. Timo und Matthias haben das Thema «Rugby» für ihren Kurzvortrag gewählt. Anhand des Vortrags beurteilen Sie heute den Wortschatz.

Machen Sie sich mit dem Ihnen ausgeteilten Beurteilungsraster zum Kriterium "Wortschatz" und der darunter stehenden Aufgabe vertraut.
Schauen Sie sich danach den Kurzvortrag an und machen Sie sich währenddessen Notizen.

Aufgabe:

Geben Sie Timo und Matthias eine kurze, konstruktive Rückmeldung zu ihrem Kurzvortrag. Stützen Sie sich dabei auf das Beurteilungsraster und Ihre Notizen.

1. Bedanken Sie sich bei Timo und Matthias.
2. Geben Sie eine kurze Rückmeldung zu einem konkreten inhaltlichen Aspekt des Kurzvortrags.
3. Geben Sie eine Rückmeldung zum Wortschatz.
4. Schliessen Sie Ihre Rückmeldung mit einem motivierenden Kommentar ab, der sich auf Timos und Matthias' Gesamtbeitrag bezieht.

Nehmen Sie Ihre Rückmeldung **auf Englisch** auf. Beachten Sie die angegebene Zeitvorgabe sowie das Niveau der Klasse. (**Sprechzeit** 1 -2 min).

Figure 29 : Sample test task 5

Zusatzmaterial Frage 5 (E/F3.2 Kurzvortrag Beurteilung)

Geben Sie Timo und Matthias eine kurze, konstruktive Rückmeldung zu ihrem Kurzvortrag. Stützen Sie sich dabei auf das Beurteilungsraster und Ihre Notizen:

Kategorie	☹	😐	☺	Kommentare
Wortschatz	❑ Die Wortwahl ist inhaltlich unpassend.	❑ Die Wortwahl ist inhaltlich grundsätzlich passend.	❑ Die Wortwahl ist inhaltlich treffend.	

Table 43 : Supplementing test material test task 5

Ihre Aufgabe für Ihre konstruktive Rückmeldung an Timo und Matthias:

1. Bedanken Sie sich bei Timo und Matthias.
2. Geben Sie eine kurze Rückmeldung zu einem konkreten inhaltlichen Aspekt des Kurzvortrags.
3. Geben Sie eine Rückmeldung zum Wortschatz.
4. Schliessen Sie Ihre Rückmeldung mit einem motivierenden Kommentar ab, der sich auf Timos und Matthias' Gesamtbeitrag bezieht.

C Profession-Related Language Competence Assessment Rubric

Allgemein: Inhaltliche Umsetzung der Aufgabe

Ausführungsniveau Qualität der Aufgabenausführung	**0**	**1**	**2**		**3**	**Bemerkungen**
Inhaltliche Umsetzung der Aufgabe	Sie/er hat **keine** inhaltlichen Vorgaben umgesetzt.	Sie/er hat **weniger als die Hälfte** der inhaltlichen Vorgaben umgesetzt. **(1-49%)**	Sie/er hat **die Hälfte oder mehr, aber nicht alle** inhaltlichen Vorgaben vollständig umgesetzt. **50-74% = 2** **75-99% = 2***		Sie/er hat **alle** inhaltlichen Vorgaben **vollständig** umgesetzt. **(100%)**	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	

Produktion: Qualitative Merkmale des Sprechens

Ausführungsniveau Qualität der Aufgabenausführung	**0**	**1**	**2**		**3**	**Bemerkungen**
Wortschatz: Wortwahl Sich im gegebenen Kontext mit inhaltlich passender Wortwahl ausdrücken	Ihre/seine Wortwahl ist im gegebenen Kontext **durchgehend inhaltlich unpassend.**	Ihre/seine Wortwahl ist im gegebenen Kontext **wiederholt inhaltlich unpassend**.	Ihre/seine Wortwahl ist im gegebenen Kontext inhaltlich **grundsätzlich passend**.		Ihre/seine Wortwahl ist im gegebenen Kontext inhaltlich **differenziert** und **treffend**.	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	
Sprachliche Korrektheit Sich sprachlich korrekt ausdrücken (Grammatik)	Sie/er macht **so häufig** grammatische Fehler, dass **durchgehend unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **häufig** grammatische Fehler, wobei **teilweise unklar** ist, was sie/er ausdrücken möchte.	Sie/er macht **manchmal** grammatische Fehler, wobei **grundsätzlich klar** ist, was sie/er ausdrücken möchte.		Sie/er macht nur **sehr selten** oder **gar nie** grammatische Fehler, die auffallen.	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	
Aussprache & Betonung Sich mit korrekter Aussprache und Betonung ausdrücken	Sie/er spricht mit einer **unverständlichen** und **unklaren** Aussprache und Betonung, was das **Verständnis durchgehend stark einschränkt**.	Sie/er spricht **wiederholt** etwas **falsch** aus oder betont etwas falsch, was zu **Verständnisproblemen** führen kann.	Sie/er spricht nur **selten** etwas **falsch** aus oder betont etwas falsch. **Grundsätzlich** ist **klar**, was sie/er ausdrücken möchte.		Sie/er spricht mit einer **gut verständlichen** und klaren Aussprache und **präzisen** Betonung (auch wenn sie/er mit einem fremdsprachlichen Akzent spricht).	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	

Flüssigkeit Sich flüssig ausdrücken, ohne zu lange oder zu viele Pausen oder Strategien zur Pausen-überbrückung einzusetzen	**Sprachliche Unsicherheiten** schränken sie/ihn so stark ein, dass **kein Redefluss zustande** kommt.	Sie/er spricht **aufgrund sprachlicher Unsicherheiten auffallend langsam** und/oder macht **oft** Pausen, um nach Ausdrücken zu suchen oder neu anzusetzen.	Sie/er spricht **aufgrund sprachlicher Unsicherheiten** mit **auffallenden Veränderungen** im Sprechtempo und/oder **gelegentlichen** Pausen.		Sie/er macht nur **selten** oder **gar keine** Pausen wegen einer **sprachlichen Unsicherheit.**	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	
Kohäsion & Kohärenz Sich sprachlich und inhaltlich zusammenhängend und strukturiert ausdrücken	Sie/er drückt sich **durchgehend nicht zusammenhängend** und **nicht klar strukturiert** aus. Allfällige sprachliche Mittel zur Verknüpfung der Äußerungen sind **unpassend**.	Sie/er drückt sich **gelegentlich nicht zusammenhängend** und **nicht klar strukturiert** aus. Sie/er verknüpft ihre/seine Äußerungen nur mit einigen wenigen sprachlichen Mitteln, die **teilweise unpassend** sind.	Sie/er drückt sich **grundsätzlich zusammenhängend** und **strukturiert** aus. Sie/er verknüpft ihre/seine Äußerungen mit einer **begrenzten** Anzahl von **passenden** sprachlichen Mitteln.		Sie/er drückt sich **durchgehend zusammenhängend** und **klar strukturiert** aus. Sie/er verknüpft ihre/seine Äußerungen **flexibel** und sicher mit **präzisen und passenden** sprachlichen Mitteln.	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	
Adressatenbezug: Lernende Sich den Lernenden gegenüber verständlich ausdrücken	Ihr/ihm **gelingt es nicht**, die Sprache an die Lernenden anzupassen, um ihnen das Verständnis zu ermöglichen.	Ihr/ihm **gelingt es nur teilweise**, die Sprache an die Lernenden anzupassen, um ihnen das Verständnis zu ermöglichen.	Ihr/ihm **gelingt es grundsätzlich**, die Sprache an die Lernenden anzupassen, um das Verständnis zu ermöglichen.		Ihr/ihm **gelingt es gut**, die Sprache an die Lernenden anzupassen, um das Verständnis zu ermöglichen.	
	trifft zu **0**	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**	

D
Excerpt Rating Manual

Beurteilung: Ausführungsniveaus

Für jede Beurteilungskomponente (z.B. «Sprachliche Korrektheit» oder «Aussprache & Betonung») sind in den Skalen vier ausformulierte Ausführungsniveaus ausgeführt (0 – 3). Bei der Beschreibung der Ausführungsniveaus wurde darauf geachtet, dass die Unterschiede zwischen den Niveaus möglichst klar erkennbar sind. Entsprechend ist bei der Beurteilung wichtig, die jeweils angrenzenden Ausführungsniveaus als Orientierungshilfe zu benutzen.

Für die Beurteilung eines sprachlichen Produktes werden die Beschreibungen zu den Ausführungsniveaus 0 bis 3 der ausgewählten Beurteilungskomponente genau durchgelesen und jene Formulierung angekreuzt, die am besten zur *sprachlichen Qualität des Produktes* passt. Die Ausführungsniveaus beziehen sich auf das *Qualitätsniveau der konkreten Aufgabenausführung* (Performanz), geben entsprechend einen Eindruck der Ausführung der spezifischen Aufgabe und stellen noch keine generalisierenden Aussagen über das allgemeine Kompetenzniveau einer Person dar.

Zusätzlich kann eine Beurteilung des Qualitätsniveaus der konkreten Aufgabenausführung durch Ankreuzen von «trifft eher zu» oder «trifft zu» auf dem Ausführungsniveau 2 feinmaschiger abgestuft werden […].

Resultate aus ersten Pilotierungen zeigen, dass die bisher erhobenen Aufgabenlösungen oft auf das Ausführungsniveau 2 zutreffen. Die Ausdifferenzierung des Ausführungsniveaus 2 dient demnach einerseits dazu, eine feinere Differenzierung bei der Beurteilung vorzunehmen und somit Mitteneffekte zu vermeiden. Andererseits soll die Ausdifferenzierung der Komplexität der Testantworten sowie dem Leistungsspektrum gerecht werden – denn ist es eher die Norm als die Ausnahme, dass eine Stufenbeschreibung komplett auf eine Aufgabenlösung passt bzw. dass eine Aufgabenlösung repräsentativ für eine einzelne Stufenbeschreibung ist.

- «Trifft eher zu» (2) wird ausgewählt, wenn die Aufgabenausführung zwar *grundsätzlich* auf die Beschreibung des gewählten Ausführungsniveaus passt, jedoch innerhalb des Niveaus noch eingeschränkt ist, relativiert werden muss und noch deutlich verbessert werden kann. Z.B. wird «Trifft eher zu (2)» bei der Komponente «Sprachliche Korrektheit» (s. Tabelle 1) gewählt, wenn entweder:
    - grammatische Fehler relativ häufig auftauchen, jedoch trotzdem grundsätzlich klar ist, was sie/er ausdrücken möchte,
    - grammatische Fehler nur manchmal auftauchen, jedoch aufgrund der grammatischen Fehler teilweise unklar ist, was sie/er ausdrücken möchte.

- «Trifft zu» (2*) wird ausgewählt, wenn die Beschreibung des Ausführungsniveaus mit dem Qualitätsniveau der konkreten Aufgabenausführung für eine Beurteilungskomponente *ohne Einschränkung* übereinstimmt.

Vorgehen: Beurteilungspakete

Bei der Beurteilung von mehreren aufeinanderfolgenden sprachlichen Produktionen wird empfohlen, jeweils eine Produktion bzw. Aufgabenlösung nach der anderen als Eigenleistung anhand aller Beurteilungskomponenten zu beurteilen. Um einer kognitiven Überlastung entgegenzuwirken wird zudem empfohlen, «Pakete» von jeweils fünf bis maximal sechs Produktionen für die Beurteilung zusammenzunehmen und auszuwerten und danach eine kurze Pause einzulegen. Anschliessend kann das nächste Paket beurteilt werden. Im Folgenden werden einige Hinweise für die Beurteilung der einzelnen Komponenten aufgeführt.

Allgemein: Inhaltliche Umsetzung der Aufgabe

Diese Skala enthält die Komponente «Inhaltliche Umsetzung der Aufgabe» und stellt die vollständige und funktional-pragmatische Umsetzung der geforderten Sprachhandlung ins Zentrum, unabhängig davon, ob die Umsetzung z.B. sprachlich korrekt oder adressatengerecht ist. Tabelle 2 enthält die Beschreibung der unterschiedlichen Ausführungsniveaus:

Inhaltliche Umsetzung der Aufgabe	Sie/er hat **keine** inhaltlichen Vorgaben umgesetzt.	Sie/er hat **weniger als die Hälfte** der inhaltlichen Vorgaben umgesetzt. **(1-49%)**	Sie/er hat **die Hälfte oder mehr, aber nicht alle** inhaltlichen Vorgaben vollständig umgesetzt. **50-74% = 2**	**75-99% = 2***	Sie/er hat **alle** inhaltlichen Vorgaben **vollständig** umgesetzt. **(100%)**
	0	trifft zu **1**	trifft eher zu **2**	trifft zu **2***	trifft zu **3**

Table 44 : Bereich «Allgemein: Inhaltliche Umsetzung der Aufgabe»

Da sich die verschiedenen Testaufgaben in der jeweiligen Anzahl der inhaltlichen Vorgaben unterscheiden, wird anhand der prozentualen Erfüllung der Vorgaben entschieden, welchem Ausführungsniveau die Aufgabenlösung zugeordnet wird (0%=0, 1-49%=1, 50-74%=2, 75-99%=2*, 100%=3). Im Kapitel 5 wird dieses Vorgehen spezifisch für jede Aufgabe genau beschrieben.

Bei der Bewertung der Komponente «Inhaltliche Umsetzung der Aufgabe» sind folgende Punkte zu beachten:

- Die Reihenfolge, in welcher die inhaltlichen Vorgaben ausgeführt werden, wird bei der Beurteilung dieser Komponente nicht berücksichtigt.
- Wenn die Aufgabe von den Testteilnehmenden nicht richtig verstanden und entsprechend falsch gelöst wird, wird das in dieser Komponente beurteilt.
- Das Einhalten der zeitlichen Richtwerte pro Aufgabe wird nicht bewertet, weil es sich hierbei nicht um eine Vorgabe, sondern um einen Anhaltspunkt bzw. eine Hilfestellung für die Testteilnehmenden bei der Testausführung handelt. Sollte eine Aufgabenlösung deutlich über oder unter dem zeitlichen Richtwert liegen, wird dies folgendermassen im Kommentarfeld vermerkt:
    - *L*: 4 Minuten oder mehr
    - *K*: 20 Sekunden oder weniger

- Sofern alle inhaltlichen Vorgaben umgesetzt werden, fliessen inhaltliche Ausführungen, die zusätzlich zu den inhaltlichen Vorgaben vorgenommen werden (z.B. inhaltliches Ausholen oder thematisches Abschweifen), *nicht* in die Bewertung dieser Komponente ein. Vielmehr werden sie als Zusatzinformationen verstanden, welche der Gestaltung des Fremdsprachenunterrichts dienen und als Teil einer authentischen Sprachhandlung im Klassenzimmer interpretiert werden können.
- Werden inhaltliche Vorgaben zwar erfüllt aber nur vage dargestellt bzw. angedeutet und nicht explizit ausformuliert, fällt die Bewertung der Aufgabenlösung in das Ausführungsniveau 2 (2 oder 2*).
- Diese Komponente beurteilt explizit nur das Erfüllen der eigentlichen Aufgabe. Umfasst eine Aufgabenlösung inhaltliche Fehlinformationen wird bei der Beurteilung folgendermassen vorgegangen:
    - Handelt es sich bei den Fehlinformationen um Allgemeinwissen bzw. Weltwissen, erfolgen keine Abzüge: Weltwissen (inkl. metalinguistisches Wissen, cf. Aufgabe 4, z.B. anstatt «Hindi» wird «Hindu» als Bezeichnung einer der indischen Landessprachen gewählt) und allgemeines Wissen werden nicht beurteilt.
    - Handelt es sich bei den Fehlinformationen um metasprachliches Wissen, kommt es in der Komponente «Sprachliche Korrektheit» zu Abzug (z.B. wird der Begriff «Syntax» falsch verwendet oder mit «Semantik» verwechselt).
    - Handelt es sich bei den Fehlinformationen um Wissen, welches sich auf die Zielsprache bezieht (z.B. Informationen zu Grammatik, Aussprache, Wortwahl etc.), wird es in der entsprechenden Komponenten beurteilt. Erklärt beispielsweise ein*e Testteilnehmende*r eine Grammatikregel falsch, erfolgen Abzüge in der Komponente «sprachliche Korrektheit».

Qualitative Merkmale des Sprechens

Diese Skala dient der Beurteilung von qualitativen Merkmalen des Sprechens wie «Sprachliche Korrektheit», «Flüssigkeit» oder «Adressatenbezug: Lernende». Dabei soll darauf geachtet werden, dass die Beurteilung der einzelnen Komponenten bei den jeweiligen Einzelleistungen unabhängig voneinander geschieht.

Allgemeine Bemerkungen

In den Beschreibungen einiger Ausführungsniveaus werden «sprachliche Mittel» erwähnt. Diese umfassen «Kenntnisse und Beherrschung der Grammatik, des Wortschatzes und der Phonologie […], die erforderlich sind, um [eine] Aufgabe auszuführen» (Council of Europe, 2001). Bei der Beurteilung wird demnach eruiert, inwiefern die sprachlichen Mittel wie der Wortschatz, grammatische Korrektheit, Flüssigkeit, Kohärenz und Angemessenheit der Sprachproduktion ausreichen, um eine gegebene Aufgabe zu erfüllen.

Bei Komponenten wie «Wortschatz: Wortwahl» und «Adressatenbezug: Lernende» ist darauf zu achten, dass sich das, was beurteilt werden soll, vor allem bei kürzeren Turns stark überschneiden kann. Bei der Beurteilung ist daher besonders wichtig zu versuchen, die unterschiedlichen Komponenten inhaltlich getrennt voneinander zu beurteilen. So soll beispielsweise die Komponente «Wortschatz: Wortwahl» entkoppelt von «Adressatenbezug: Lernende» und rein anhand der sprachlichen Qualität von «Wortschatz» (z.B. Wortschatztiefe und -breite, Akkuratheit etc.) beurteilt werden. Die Fähigkeit, sich der Zielstufe angepasst auszudrücken wird demnach lediglich bei der Komponente «Adressatenbezug: Lernende» beurteilt.

Wortschatz: Wortwahl

Wie erwähnt, wird diese Komponente entkoppelt vom Adressatenbezug beurteilt. Folgende Aspekte gilt es bei dieser Komponente zu beachten:

- Eine Aufgabenlösung wird bei Ausführungsniveau 1 eingestuft, wenn wiederholt inhaltlich unpassende Wörter gewählt werden. Das Verständnis der Hauptaussage der Aufgabenlösung ist aufgrund der falschen Wortwahl beeinträchtigt und nicht sichergestellt. Dies kann sich entweder darin niederschlagen, dass ein einzelnes für das Verständnis der Aussage notwendiges Wort semantisch falsch ist, unpassend gewählt und wiederholt unpassend eingesetzt wird, oder dass mehrere unpassende Wörter, welche zwar das Verständnis weniger stark beeinträchtigen, wiederholt falsch eingesetzt werden (= ein essenzielles Wort oder mehrere semi-essenzielle Wörter wiederholt falsch einsetzen).
- Repetitive Wortwahl, auch wenn inhaltlich korrekt, wird hier als Hinweis auf eingeschränkte Wortschatztiefe und -breite interpretiert.
- Die Beurteilung von Wortschatztiefe und -breite schlägt sich in den Ausführungsniveaus 2, 2* und 3 und den Adjektiven *passend* und *treffend* nieder:
    - 2/2*: «*Ihre/seine Wortwahl ist im gegebenen Kontext inhaltlich grundsätzlich passend*»:
        - Allgemein: Unter *inhaltlich passend* wird verstanden, dass die gewählten Worte generell thematisch passend ausgewählt wurden, jedoch noch vage und eingeschränkt differenziert sind (Beispiel: «things»).
        - Eine Aufgabenlösung wird im Ausführungsniveau 2 eingestuft, wenn Ambiguitäten mit hoher Wahrscheinlichkeit auftauchen und kleinere Missverständnisse daraus resultieren können. Einzelne Wörter können inhaltlich unpassend (z.B. wenn «but» anstatt «and» verwendet wird) und das Verständnis der Hauptaussage leicht beeinträchtigt sein.
        - Eine Aufgabenlösung wird dann dem Niveau 2* zugeteilt, wenn die Wortwahl grundsätzlich inhaltlich passend und die Hauptaussage der Aufgabenlösung grundsätzlich verständlich ist. Leichte Ungenauigkeiten können noch auftauchen.
    - 3: «*Ihre/seine Wortwahl ist im gegebenen Kontext inhaltlich differenziert und treffend*»: Unter *inhaltlich differenziert und treffend* wird verstanden, dass die gewählten Worte inhaltlich präzise sind und den Inhalt klar und prägnant wiedergeben. Eine inhaltlich differenzierte und treffende Wortwahl lässt keine Ambiguitäten zu (Beispiel: «factors», cf. «things» oben).
- Umgangssprachliche Ausdrücke wie gängige und gut verständliche Kurzformen bzw. stilistische Abkürzungen (z.B. «gonna»), Füllwörter und Phrasierungen (z.B. «like», «you know» etc.) oder Unterschiede zwischen Standard und non-Standard Varietäten (z.B. «doing» vs «doin'») gelten hier als Merkmale der gesprochenen Sprache. Treten solche Ausdrücke auf, sollten sie im Kommentarfeld vermerkt werden. Diese Merkmale werden demnach ohne Konsequenzen auf die Beurteilung dieser Komponente akzeptiert. Schränken umgangssprachliche Ausdrücke das Verständnis der Lernenden ein, werden sie in der Komponente «Adressatenbezug: Lernende» beurteilt (cf. 5.4.7).
- Von den Testteilnehmenden wiedergegebene Fehlinformationen bezüglich Wortschatz deuten auf inkorrektes Wortschatzwissen hin und werden entsprechend in dieser Komponente beurteilt (z.B. «*Penalty* is not an English word and should not be used»).

Sprachliche Korrektheit

Diese Komponente dient der Beurteilung der sprachlichen Korrektheit entkoppelt von jeglichen anderen Komponenten. Bei der Beurteilung von Ungenauigkeiten und Fehlern wird nicht zwischen schwerwiegenderen und weniger schwerwiegenden Fehlern unterschieden. Zu beachten ist hier die relative Häufigkeit von Fehlern in Bezug auf die aktive Sprechzeit. Treten in einer Aufgabenlösung mit kurzer Sprechzeit gleich viele Fehler auf wie in einer Aufgabenlösung mit langer Sprechzeit, wird erstere tiefer beurteilt als letztere. Diese Einschätzung beruht jeweils auf dem Eindruck der jeweiligen Aufgabenlösung.

Von den Testteilnehmenden wiedergegebene Fehlinformationen bezüglich Aspekten wie Grammatik, Syntax etc., welche die sprachliche Korrektheit thematisieren, deuten auf inkorrektes Sprachwissen hin und werden entsprechend in dieser Komponente beurteilt (z.B. ein*e Testteilnehmende*r weist eine*n Schüler*in fälschlicherweise darauf hin, das Present Perfect anstatt das Present Continuous zu benutzen).

Aussprache und Betonung

In dieser Komponente wird die Korrektheit von Aussprache und Betonung beurteilt. Analog zu den neuen Deskriptoren des Companion Volume des Gemeinsamen Europäischen Referenzrahmens (Council of Europe, 2018) gilt hier «the ideal native speaker» nicht als die höchste und zu erreichende Stufe der Sprachkompetenz. Entsprechend führen starke fremdsprachliche bzw. muttersprachliche Akzente nicht zu Abwertungen bei der Leistungsbeurteilung – es sei denn, es handelte sich um Aussprachefehler wie einen sinnentstellenden Wortakzent, der leicht zu Missverständnissen führen kann, etwa im Falle von Homografen wie /ˈmɒd(ə)n/ versus /məˈdəːn/ (Arras, 2011; Europe, 2018) oder /dɪˈvɛləp/ versus /ˈdɛvɛləp/. Wird hingegen «interesting» mit leicht verschobener Betonung ausgesprochen (/ɪnt(ə)rɪˈstɪŋ/ anstatt /ˈɪnt(ə)rɪstɪŋ/), führt dies zu keinem Abzug. Grundsätzlich gilt - gemäss der Kompetenzorientierung des GER – dass die Verständigung bzw. die Kommunikation bei der Beurteilung im Mittelpunkt stehen, «die trotz eines Akzents oder bestimmter Intonationsfehler durchaus gewährleistet sein kann» (Arras, 2011).

Flüssigkeit

Bei der Beurteilung der Komponente «Flüssigkeit» sind folgende Hinweise zu beachten:

- Pausen, die dazu genutzt werden, um nach Wörtern oder Formulierungen zu suchen, werden zur Beurteilung herangezogen. Hingegen Pausen, die offensichtlich didaktisch motiviert sind (z.B. Pausen, die den Lernenden ermöglichen, das Gehörte zu verarbeiten), werden bei der Beurteilung nicht penalisiert. Solche didaktischen Pausen sollten in der Regel durch die Produktion unterscheidbar sein von Pausen, die aufgrund von sprachlicher Unsicherheit zum Finden eines Wortes oder einer Formulierung eingesetzt werden. Ist sich ein*e Beurteilende*r bei den hörbaren Pausen unsicher, um welche Art von Pausen es sich bei der konkreten Aufgabenlösung handelt, soll dies im Kommentarfeld entsprechend vermerkt werden.
- Auch hier gilt es, Bewertungen, die sich aufgrund von Pausen ergeben, nicht doppelt vorzunehmen. So sollen beispielsweise Pausen nur bei der Komponente «Flüssigkeit» und nicht zusätzlich auf dem Ausführungsniveau 1 bei «Wortschatz: Wortwahl» (doppelt) beurteilt werden.
- Tempo: Analog zu den anderen Beurteilungskomponenten wird auch «Flüssigkeit» (hier am Beispiel des konkreten Aspekts «Sprechgeschwindigkeit») unabhängig vom Adressatenbezug beurteilt. Zu unterscheiden gilt es bezüglich Sprechgeschwindigkeit zwischen Folgendem:
    - Spricht ein*e Testteilnehmende*r *zu schnell* für die Zielstufe, wird dies in der Komponente «Adressatenbezug: Lernende» berücksichtigt und beurteilt. Eine

zu hohe Sprechgeschwindigkeit wird hier nicht als Merkmal von sprachlicher Unsicherheit interpretiert, sondern als eingeschränkte Fähigkeit, sich sprachlich dem Niveau der Zielstufe anzupassen.
    - Spricht ein*e Testteilnehmende*r aufgrund sprachlicher Unsicherheiten auffallend langsam, wird dies in der Komponente «Flüssigkeit» beurteilt.
    - Einzuschätzen, ob die Pausen aufgrund sprachlicher Unsicherheiten entstehen kann stark auf Vermutungen der Beurteilenden basieren und entsprechend eine jeweils hypothetische Dimension umfassen. Unsicherheiten sollten entsprechend im Kommentarfeld vermerkt werden.
- Die Lautstärke einer Sprachproduktion ist in diesem Kontext oft situationsbedingt und wird durch die Testsituation beeinflusst. Aus diesem Grund wird die Lautstärke nicht beurteilt bzw. fliesst nicht in die Beurteilung der «Flüssigkeit» ein.
- Füllwörter wie «ähm», «so», «well», «like» etc. gelten als Merkmale der gesprochenen Sprache. Werden auffallend viele Füllwörter eingesetzt und kann dadurch auf sprachliche Unsicherheiten gedeutet werden, fliesst das in Beurteilung der «Flüssigkeit» ein (z.B. wenn Füllwörter als Strategie verwendet werden, um sprachliche Unsicherheiten zu überdecken und gleichzeitig eine hohe Flüssigkeit beizubehalten).
- Weitere Strategien wie z.B. Wiederholungen (lexikalisch und argumentativ), schnelles Neu-Ansetzen (ohne merkliche Pause), kurzes Zögern, unvollständige Sätze etc., solange «natürlich», werden nicht penalisiert sondern gelten als Merkmale der Mündlichkeit.

Kohäsion & Kohärenz

Diese Komponente erfasst sowohl die sprachliche als auch inhaltliche Kohäsion und Kohärenz. Abzug bei dieser Komponente erfolgt:

- wenn sich eine Aufgabenlösung inhaltlich im Kreise dreht (z.B. die Realisierung einer Aufgabe bzw. die Argumentation ist inhaltlich unnötig repetitiv oder ein*e Testteilnehmende*r kommt inhaltlich nicht auf den Punkt),
- aufgrund eingeschränkter Verlinkung der Inhalte die Nachvollziehbarkeit der Äusserungen erschwert ist,
- der rote Faden der Aufgabenlösung nicht erkennbar ist.

Folgende konkrete Punkte gilt es demnach bei der Beurteilung zu beachten:

- Ist eine Aufgabenlösung inhaltlich und sprachlich nicht (immer) klar strukturiert und es werden unpassende oder nur einfache / begrenzte sprachliche Mittel eingesetzt, dann fällt das in das Ausführungsniveau 1.
- Ausführungsniveau 2 ist erreicht, wenn eine Aufgabenlösung inhaltlich und sprachlich nicht (immer) klar strukturiert ist und die eingesetzten sprachlichen Mittel begrenzt sind. Demnach werden keine oder kaum sprachliche Mittel wie Konnektoren eingesetzt, und jene, die eingesetzt werden, sind entweder repetitive oder sehr simpel, z.B. «and», «so», «but» etc.)
- Eine Aufgabenlösung fällt ins Ausführungsniveau 2* wenn:
    - eine Aufgabenlösung inhaltlich und sprachlich nicht (immer) klar strukturiert ist, aber differenzierte sprachliche Mittel eingesetzt werden (z.B. «however», «first of all», «finally» etc.)
    - eine sowohl sprachlich als auch inhaltlich klare und zusammenhängende Struktur vorhanden ist – der rote Faden ist klar erkennbar – diese jedoch mit sehr einfachen sprachlichen Mittel hergestellt wird (e.g., repetitive «so», «and», «but» etc.)

- Ist eine sowohl inhaltlich als auch sprachlich klare Struktur und zusammenhängende Darstellung vorhanden, und Kohäsion und Kohärenz wird mit differenzierten sprachlichen Mitteln hergestellt (e.g., «however», «furthermore», «moreover», «in addition», «nevertheless» etc.), fällt die Aufgabenlösung ins Ausführungsniveau 3.
- Zudem gilt es sich des inhaltlichen Wiederspruchs dieser Komponente bewusst zu sein: Bei der Beurteilung muss einerseits auf Details wie den Einsatz von sprachlichen Mitteln (quantifizierbar) geachtet und andererseits das Gesamtbild bzw. den Gesamteindruck der Aufgabenlösung hinsichtlich Kohäsion und Kohärenz (nicht quantifizierbar) in Betracht gezogen werden (kann man der Aussage gut folgen?). Diese Diskrepanz in einer Komponente zu vereinen ist eine Herausforderung – und trotzdem sind beide Aspekte hier ausschlaggebend.

Zusammenfassend bedeutet das:

- Ausführungsniveau 1) ist erreicht, wenn a) UND b) zutreffen
    a) Aufgabenlösung ist sprachlich nicht (oder kaum) klar strukturiert
    b) Es werden keine oder teilweise unpassende sprachliche Mittel eingesetzt
- Ausführungsniveau 2) ist erreicht, wenn a) UND b) zutreffen, ODER c) UND d) zutreffen
    a) Aufgabenlösung ist sprachlich nicht (oder kaum) klar strukturiert
    b) Es werden eine begrenzte Anzahl an passenden sprachlichen Mittel eingesetzt, die repetitive oder simpel sind
    ODER
    c) Aufgabenlösung ist grundsätzlich zusammenhängend und strukturiert
    d) Es werden keine oder teilweise unpassende sprachliche Mittel eingesetzt
- Ausführungsniveau 2*) ist erreicht, wenn a) UND b) zutreffen
    a) Aufgabenlösung ist grundsätzlich zusammenhängend und strukturiert
    b) Es werden eine begrenzte Anzahl an passenden sprachlichen Mittel eingesetzt, die tendenziell eher simpel sind.

Adressatenbezug: Lernende

Die Beurteilung der Komponente «Adressatenbezug: Lernende» ist mitunter eine der anspruchsvollsten Aufgaben im Beurteilungsprozess der mündlichen, berufsspezifischen Sprachkompetenzen, weil auch hier eine hypothetische Dimension bei der Einschätzung stark zum Tragen kommen kann. So müssen von den Beurteilenden jeweils Vermutungen angestellt werden, ob die hörbaren Äusserungen von hypothetischen Lernende der Zielstufe verstanden würden oder nicht.

Bei dieser Komponente liegt der Hauptfokus auf der Bewertung der Komplexität der Sprachproduktion. Folgende Punkte sollen eine akkurate Einschätzung erleichtern:

- Vor der Beurteilung ist es hilfreich, die Beschreibung der Klasse bzw. Lernenden bei der jeweiligen Aufgabe nochmals durchzulesen, um die Beurteilung für sich zu kontextualisieren.
- Eine klare und präzise Unterscheidung der Niveaus «Grundanforderungen» und «erweiterte Anforderungen» bei der Beurteilung des «Adressatenbezugs: Lernende» vornehmen zu können ist anspruchsvoll, zumal sich nicht nur Klassen auf gleicher Niveaustufe beträchtlich voneinander unterscheiden können sondern vielmehr auch die Profile der einzelnen Schüler*innen innerhalb der Klassen. Zudem hängen die Kompetenzen der Schüler*innen von vielen externen und internen Faktoren ab, die nicht kontrollierbar sind. Der Anspruch ist demnach vielmehr, ein angemessenes

Globalurteil mit Bezug zur Alters- und Zielstufe zu fällen, ohne sich in den detaillierten und teilweise unvergleichbaren Unterschieden der Niveaustufen zu verlieren.

- Die unten aufgeführten Aspekte können als Merkmale nicht-adressatengerechter gesprochener Sprache verstanden werden, auf welche bei der Beurteilung besonders geachtet werden soll. Kommen folgende Merkmale in Aufgabenlösungen vor, muss jeweils abgewogen werden, inwiefern sie für die Lernenden auf der Zielstufe *zu anspruchsvoll* sind und Verständnisprobleme verursachen können. Sollte es aufgrund der folgenden Merkmale zu einem Abzug in der Bewertung dieser Komponente kommen, wird dies entsprechend im Kommentarfeld vermerkt.
    - Hohe Sprechgeschwindigkeit
    - Slang (z.B. «chick», «sweet as», «chap», «lit», «on fleek» etc.)
    - Jargon (z.B. rechtswissenschaftliche Ausdrücke wie «ad hoc», IT-spezifische Ausdrücke wie «cache», Ausdrücke aus den Naturwissenschaften wie «gluteus maximus» etc.)
    - Idiomatische Ausdrücke (z.B. «dig the well before you're thirsty», «kick the bucket», «you hit the nail on the head»)
    - Partikelverben («phrasal verbs» wie z.B. «set out», «take off», «put up with»)
    - Komplexe Wortwahl (z.B. akademische, fachwissenschaftliche oder domänenspezifische Ausdrücke, niederfrequente Wörter etc.)
    - Komplexe Syntax (z.B. Sprachproduktionen, die näher an der Schriftlichkeit als Mündlichkeit liegen)
    - Nicht-Standardsprache bzw. Varietäten: Im Unterschied zur Komponente «Wortschatz: Wortwahl», wo Abweichungen von Standardvarietäten *nicht* bewertet werden, fliessen solche Abweichungen hingegen in die Bewertung von «Adressatenbezug: Lernende» ein. Somit werden eine doppelte Bewertung bzw. ein doppelter Abzug solcher Abweichungen vermieden.
- Enthält eine Sprachproduktion Aspekte wie didaktisch und inhaltlich nicht adressatengerechte Ausführungen (z.B. eine Rückmeldung ist stark mängelorientiert formuliert oder kognitiv überfordernd, weil sie ein grosses Mass an Inhalten enthält), werden diese nicht für die Beurteilung berücksichtigt. Analog zu den anderen Beurteilungskomponenten fliesst didaktisches Wissen, allgemeines Wissen oder Weltwissen nicht in die Beurteilung mit ein, weil diese Faktoren keinen Bezug zu den zu testenden Sprachkompetenzen aufweisen.

Annotierte Musterlösungen / Benchmarks

Allgemeine Bemerkungen zu den Testaufgaben

Beim für diese Dissertation entwickelten Test handelt es sich um einen kompetenzorientierten Performanztest, in welchem mündliche berufsspezifische Sprachproduktionen anhand von realweltlichen Stimuli elizitiert werden. Die Testsituation beinhaltet, dass die Testteilnehmenden unter Aufsicht an einem Rechner den Test ausführen. Dass solche Umstände Effekte auf die dargelegten Performanzen hat ist anzunehmen. Entsprechend ist einerseits wichtig zu beachten, dass bei der Testdurchführung ein Beobachtungsfaktor in die Aufgabenlösungen mit einfliesst. Andererseits sind sich angehende Lehrpersonen Beobachtungssituationen ausbildungsbedingt gewohnt (man kontrastiere beispielsweise mit dem Berufsalltag einer Reinigungskraft): Praktika sind ein gutes Beispiel dafür. Entsprechend ist ein allfälliger Beobachtungsfaktor zwar zu beachten und bei der Beurteilung von Aufgabenlösungen im Hinterkopf zu behalten (cf. z.B. „Flüssigkeit“ und leises Sprechen), ist methodisch jedoch vertretbar und erklärbar.

Die einzelnen Testaufgaben enthalten klare und systematisch dargelegte Instruktionen. Diese geben transparente Strukturen vor, anhand welcher die Sprachkompetenz von Testteilnehmenden erhoben werden soll. Diese Instruktionen sind während dem Ausführen des Tests nicht zwingend ständig sichtbar. Entsprechend gestaltet sich das Lösen einer Aufgabe als kognitiv anspruchsvoll, da die Testteilnehmenden mehrere Aspekte gleichzeitig bearbeiten müssen (z.B., sich den Kontext der Aufgabe, den Inhalt der Videovignette und die inhaltlichen Vorgaben merken, eine angemessene Antwort in der Fremdsprache formulieren etc.). Obwohl die Testsituation nicht flächendeckend mit einer realen Klassenzimmersituation verglichen werden kann, stellen die Testaufgaben doch eine Annäherung an Szenarien dar, die im authentischen Klassenzimmer vorkommen können.

Des Weiteren ist zu beachten, dass es sich beim Prä- und Post-Test zwei Mal um den gleichen Test mit den gleichen Testaufgaben handelt. Die Testteilnehmenden wurden sowohl zur Vorbereitung auf den Prä- als auch auf den Post-Test mit dem Testformat vertraut gemacht. Zwischen den Messzeitpunkten des Prä-Tests (T0) und Post-Tests (T1) liegt ein Jahr. Trotz der Zeitspanne kann ein gewisser Trainingseffekt nicht komplett ausgeschlossen werden, besonders hinsichtlich der Vertrautheit mit dem Testformat.

Aufgabe 5

Hinweise: Eine Rückmeldung zum Wortschatz ist ausreichend, um Aufgabenpunkt 3 erfolgreich zu bearbeiten. Wenn zusätzlich eine Rückmeldung zum freien Sprechen genannt wird, hat dies keinen Einfluss auf die Beurteilung (cf. „Inhaltliche Umsetzung der Aufgabe).

Beurteilungsraster

Richtzeit Aufgabe: ca. 4 min / **Umfang Antwort:** 1-2 min Sprechzeit / **Zielgruppe:** 3. Klasse Oberstufe, Realschule (allgemeine Anforderungen)

Situation:

Im Englischunterricht behandeln Sie zur Zeit das Thema «Britische Kultur» mit speziellem Fokus auf den dazugehörigen Wortschatz. Für eine summative Prüfung haben Sie Ihre Schüler*innen beauftragt, in Zweiergruppen Kurzvorträge zu einem Teilaspekt des Themas zu halten. Timo und Matthias haben das Thema «Rugby» für ihren Kurzvortrag gewählt. Anhand des Vortrags beurteilen Sie heute den Wortschatz.

Machen Sie sich mit dem Ihnen ausgeteilten Beurteilungsraster zum Kriterium "Wortschatz" und der darunter stehenden Aufgabe vertraut. Schauen Sie sich danach den Kurzvortrag an und machen Sie sich währenddessen Notizen.

Aufgabe:

Geben Sie Timo und Matthias eine kurze, konstruktive Rückmeldung zu ihrem Kurzvortrag. Stützen Sie sich dabei auf das Beurteilungsraster und Ihre Notizen.

1. Bedanken Sie sich bei Timo und Matthias.
2. Geben Sie eine kurze Rückmeldung zu einem konkreten inhaltlichen Aspekt des Kurzvortrags.
3. Geben Sie eine Rückmeldung zum Wortschatz.
4. Schliessen Sie Ihre Rückmeldung mit einem motivierenden Kommentar ab, der sich auf Timos und Matthias' Gesamtbeitrag bezieht.

Nehmen Sie Ihre Rückmeldung **auf Englisch** auf. Beachten Sie die angegebene Zeitvorgabe sowie das Niveau der Klasse. (**Sprechzeit** 1 -2 min).

Figure 30 : Benchmark test task 5

Kategorie	☹	😐	☺	**Kommentare**
Wortschatz	❑ Die Wortwahl ist inhaltlich **unpassend.**	❑ Die Wortwahl ist inhaltlich **grundsätzlich passend**.	❑ Die Wortwahl ist inhaltlich **treffend.**	

Komponenten / Lösung	Inhaltl Umsetzung der Aufgabe	Wortschatz: Wortwahl	Sprachliche Korrektheit	Aussprache und Betonung	Flüssigkeit	Kohäsion und Kohärenz	Adressatenbezug: Lernende
9166_dOm_5							
Rater Training	3 Alle Vorgaben erfüllt	2* nicht immer ganz klar, aus strenger linguistischer Perspektive ist «agree wiith you» aus pragmatischer Sicht zwar falsch, ist nicht komplett treffend aber immer noch grundsätzlich passend. «Penalty is not an English word» = Fehlinformation zu Wortschatz, gibt hier Abzug.	2 «you did good», «a good vocabulary», «a short feedback», expressed yourself good», «some days». Abgrenzung zu 1: relative Häufigkeit der Fehler, aber immer noch klar was sie sagen will.	2 Manchmal starke Betonung auf gewissen Silben, kann leicht verwirren. /v/ vs /w/ «Timo», «wif», «familiar», «Rugby» (Schweizerisch ausgesprochen), L1 Einfluss grenzt ans Störende, ist grundsätzlich klar was sie sagen möchte. Spricht ein bisschen öfter als nur selten etwas falsch aus. Spricht auch wiederholt das gleiche Wort falsch aus. «tschob». Fehler sind relative offensichtlich und auffällig. Häufigkeit durch wiederholte falsche Aussprache.	3 manchmal etwas zögerlich aber nicht wahnsinnig auffällig. Zögern eher aufgrund Denkpausen und nicht sprachlicher Unsicherheiten.	3 gut verknüpft, einmal etwas unpassend aber generell passend	1 Niveau sehr komplex, zu komplex für Grundanforderungen. «Terms», «express yourself», sehr hohes Niveau obwohl simpler als vorher, macht Anstalten zum Anpassen. Tempo relativ gut angepasst. Aussprache teilweise unklar, teilweise sehr gut und klar betont. Adressatenbezug gelingt nicht gut bzw. nur teilweise, das Niveau anzupassen.
9037_dOm_5							

Rater Training	3 Am Anfang und am Schluss Rückmeldung zu Inhalt (nicht sehr kurz), die inhaltliche Rückmeldung am Schluss zu „motivierender Kommentar, der sich auf den Gesamtbeitrag bezieht“ verstanden werden kann; mehrere Rückmeldungen zu Wortschatz (ist aber OK, weil Vorgabe „eine Rückmeldung zum Wortschatz“ ist und nicht „eine Rückmeldung zu einem konkreten Aspekt des Wortschatzes“, cf. Aufgabenpunkt 1).	2 Teilweise gut und präzise wie bei „loved the fact that you explained“, dann teilweise unpassend und inkorrekt: “football form”, “synonyms for rugby” = Fehlinformation, teilweise repetitive Wortwahl (“presentation… presentations”), “different parts of football” → teilweise unpassend, aber nicht „wiederholt inhaltlich unpassend“, deshalb 2.	2 Manchmal grammatische Fehler, Syntaxprobleme treten auf sowie inaccuracies bei bswp. „more various“, “You explained us how the…», “various” vs “varied”, “for your vocabulary” etc. Grundsätzlich ist klar, was sie sagen möchte.	2 etwas öfter falsche Aussprache als «selten» wie im Deskriptor (/countries/, Probleme mit /the/, /rugby/, grundsätzlich jedoch klar was sie sagen möchte), deshalb 2 und nicht 2*	2 Relativ lange Aufnahme, teilweise viele “ähms”, auffallende Veränderungen im Sprechtempo und öfters mal Pausen (mehr als nur gelegentlich) und etwas stockend.	2 Insbesondere durch die lange, sich wiederholende Argumentation am Ende reduziert. Ein paar gute devices, e.g., „first of all“, „for example“, jedoch „so“ und „and“ etwas repetitiv. Aufgrund der Länge wird inhaltliche Struktur beeinträchtigt.	2* wiederholt lange/komplizierte Schachtelsätze, ausserdem sehr schnell gesprochene Sequenzen (und gleichzeitiges Stocken etc.), Tempo und Wortwahl grundsätzlich passend.

Table 45 : Annotated benchmark test task 5

E

Interview Guide Sub-Study

Interviewleitfaden Erhebung Schüler*innenperspektive

OZ Buechenwald, 27.10.2020

OZ Buechenwald, 3. & 4.11.2020

Forschungsfragen

- Wie schätzen Schüler*innen der Sekundarstufe 1 ein mündliches, Englisches Feedback von angehenden Fremdsprachenlehrpersonen bezüglich dessen sprachlicher Qualität und Verständlichkeit ein?
- Welche sprachlichen und inhaltlichen Aspekte werden von Schüler*innen der Sekundarstufe 1 wahrgenommen und zur Sicherstellung des Verständnisses als unabdingbar empfunden?

Vorbereitung

- Batterien des Aufnahmegeräts kontrollieren
- PC hochfahren und auf Moodle Aufgabe 3 öffnen
- Auf SWITCHdrive «Haupterhebung Aufgabenbeispiele» öffnen:
    - Aufgabe 3
        - 8769_dOm_3_low prof, low adressateng.
        - 8842_dOm_3_low prof, medium adressateng.
        - 9039_dOm_3_high prof, high adressateng.
        - 9099_dOm_3_high prof, low adressateng.
- Leitfaden öffnen und Angaben von Teilnehmenden auf Leitfaden notieren
- Interviewleitfaden nochmals durchlesen, damit der Ablauf präsent ist
- Aufnahmefunktion vorbereiten

Eröffnung Interview und Begrüssung

- Danke fürs Mitmachen bei diesem Interview. Es freut mich sehr, dass Du da bist.

- Bevor wir mit dem Interview beginnen, möchte ich Dir einige Informationen zum Inhalt und Ablauf geben.

- Wir führen dieses Interview auf Hochdeutsch durch.

- Zuerst etwas zum Hintergrund: Zukünftige Lehrpersonen, die einmal Englisch unterrichten werden, mussten bei uns einen Sprachtest machen. Durch diesen Test können wir schauen, wie gut sie Englisch können. In den Testaufgaben mussten die zukünftigen Lehrpersonen vor allem Schülerinnen und Schülern auf Englisch ein Feedback geben. Wir schauen mit Experten an der PH an, wie gut sie das gemacht haben.

- Weil es für uns sehr wichtig ist, dass die Schüler*innen das Englisch ihrer Lehrpersonen gut verstehen, möchten wir auch die Meinung dazu von Schüler*innen als Experten hören.

- Deshalb möchte ich heute gerne von Dir erfahren, wie gut Du die zukünftigen Lehrpersonen verstehst und wie gut Du ihr Englisch einschätzt.

- Das heisst, dass wir heute den Spiess umdrehen: anstatt dass Du wie sonst immer von den Lehrpersonen beurteilt wirst, darfst Du sie heute einmal beurteilen.

- Dazu schauen wir uns gemeinsam zwei bis vier Testantworten von diesen zukünftigen Lehrpersonen an.

- Wichtig ist mir Deine eigene Wahrnehmung und ganz persönliche Meinung. Es gibt keine richtigen oder falschen Antworten, da die Wahrnehmungen sehr individuell sind. Mich interessiert einfach, was Du ganz persönlich von den Aufnahmen hältst.

- Ich nehme das Interview auf, damit ich es später auswerten kann. Die Daten werden anonymisiert. Das heisst, dass am Schluss niemand ausser mir wissen wird, dass genau Du das gesagt hast, was wir heute aufnehmen.

- Das Interview dauert ca. 30 Minuten. Wenn Du Fragen hast, kannst Du mich jederzeit unterbrechen.

- Du darfst das Interview auch jederzeit und ohne Begründung abbrechen.

- Hast du noch Fragen bevor wir beginnen?

- Ich werde jetzt mit der Aufnahme starten.

➔ Aufnahme starten

Hinweis zur Verwendung des Leitfadens

1. Leitfrage stellen
2. Antwort abwarten
3. Beim Stocken der interviewten Person: Pause aushalten!
4. Geeignete Aufrechterhaltungs- oder Steuerungsfrage stellen
5. Antwort abwarten
6. Erst jetzt: Präzisierungsfragen stellen!

Mögliche Aufrechterhaltungs- und Steuerungsfragen

– Gibt es sonst noch etwas?
– Kannst du das genauer beschreiben?
– Wie meinst du das (genau)?
– Warum ist das so?
– Kannst du ein konkretes Beispiel dafür geben?
– Wie nimmst du diese Situation wahr?
– Gibt es noch weitere […z.B. Ziele, Schwierigkeiten, wichtige Erkenntnisse, Erfahrungen, etc.]?

Diese Aufrechterhaltungs- und Steuerungsfragen können situativ je nach Bedarf eingesetzt werden. Grundsätzlich sollen jedoch alle Themen abgedeckt werden, die Reihenfolge spielt dabei keine Rolle. Werden einzelne Aspekte bereits selbständig ausführlich von der interviewten Person thematisiert, können die entsprechenden Präzisierungsfragen übersprungen werden. Die Kästchen bei den Themen sind zum Abhaken behandelter Themen während des Interviews gedacht.

Interviewleitfaden

	Leitfrage / Erzählaufforderung	Themen	Präzisierungsfragen
1	**Einstieg: Dein erster Eindruck** *Normalerweise beurteilen Deine Lehrpersonen Deine Leistungen. Heute drehen wir den Spiess um und Du darfst den Lehrpersonen ein Feedback geben und sie beurteilen.*		
1.1	Bitte erzähl mir in einem ersten Schritt: Was ist Dein erster Eindruck dieses Feedbacks der Lehrperson?	☐ Grund erster Eindruck	- Weshalb hinterlässt die Lehrperson diesen Eindruck auf Dich?
1.2	Was hat die Lehrperson genau gesagt?	☐ Verständnis	- Wie würdest du die einzelnen Punkte des Feedbacks nochmals in deinen eigenen Worten zusammenfassen? - Gibt es noch weiteres, das im Feedback gesagt wurde? - Was ist Dir noch unklar? Was würdest Du rückfragen? - Möchtest du das Audio dazu nochmals anhören?
1.3	Gibt es etwas, das Du nicht verstanden hast?	☐ Inhalt ☐ Grund	- Was hast Du nicht verstanden? - Weshalb hast Du es nicht verstanden?
1.4	Erzähle mir doch gerne mal, was die Lehrperson ganz allgemein in deinen Augen gut gemacht hat.	☐ Grund	- Weshalb hast Du es so wahrgenommen?
1.5	Was hat die Lehrperson weniger gut gemacht?	☐ Grund	- Weshalb findest Du, hat das die Lehrperson weniger gut gemacht?
2	**Sprachgebrauch / Sprachkompetenz** *Jetzt interessiert mich besonders, wie gut Deiner Meinung nach das Englisch dieser Lehrperson ist.*		
2.1	Was hat Dir an der Sprache der Lehrperson gut oder weniger gut gefallen?	☐ Grund	- Weshalb hat Dir dieser Aspekt gefallen? - Weshalb hat Dir dieser Aspekt nicht gefallen?

2.2	Wie gut spricht diese Person Englisch?	☐ Wahrgenommene Sprachkompetenz ☐ Korrektheit	- Wie kompetent in der Fremdsprache Englisch wirkt diese Person auf Dich? - Woran erkennst Du, dass die Lehrperson gut/noch nicht so gut im Englisch ist?
2.3	Wie anstrengend war es, die Lehrperson sprachlich zu verstehen?	☐ Sprache	- Wo musstest Du Dich besonders anstrengen, damit Du die Lehrperson verstehen konntest?
2.4	Wie hat die Aussprache der Lehrperson für Dich geklungen?	☐ Verständlichkeit ☐ Akzent	- Wie verständlich war die Aussprache der Lehrperson für dich? - Was klang besonders gut? - Weshalb klang das gut für Dich? - Was klang weniger gut? - Weshalb klang das weniger gut für Dich?
2.5	Wie gut würde eine Person die Lehrperson verstehen, die im Englisch eher Mühe hat?	☐ Verständlichkeit	- Inwiefern wäre es hilfreich, wenn die Lehrperson ihr Englisch dem Niveau dieses Schülers anpassen würde?
2.6	Wie fandst Du den Wortschatz der Lehrperson?	☐ Wortwahl ☐ Wortschatz: Umfang ☐ Komplexität / Schwierigkeit	- Findest Du, hat die Lehrperson die Wörter passend / treffend zum Inhalt benutzt? - Denkst du, dass sie Lehrperson eher einen grösseren oder kleineren Wortschatz hat? - Wie schwierig / einfach war es für Dich, die Wörter zu verstehen?
2.7	Wie flüssig fandst Du das Englisch der Lehrperson?	☐ Stocken / Stottern / Pausen ☐ Tempo	- Woran erkennst Du, dass die Lehrperson flüssig/nicht flüssig Englisch spricht? - Fandst Du das Tempo angemessen, damit Du die Lehrperson verstehen konntest?
2.8	Woran müsste die Lehrperson noch arbeiten, um besser im Englisch zu werden?	☐ Feedback ☐ Verbesserungstipps	- Welches Feedback würdest Du der Lehrperson zu ihrem Englisch geben? - Welche Tipps würdest Du der Lehrperson zur Verbesserung ihres Englisch geben?

2.9	Was möchtest Du sonst noch zu dem sagen, was Du von der Lehrperson gehört hast?	☐ Weiteres ☐ Fragen	
3	**Feedback: Gewohnheit** *Jetzt möchte ich gerne mit dir darüber sprechen, welche Art von Feedback Du Dir in der Schule gewohnt bist und wie das, was Du gehört hast, dem entspricht oder nicht entspricht.*		
3.1	In welcher Sprache gibt Dir Deine Englischlehrperson Feedback?		
3.2	Wie unterscheidet sich dieses Feedback von demjenigen, das Du von Deiner Englischlehrpersonen kennst?	☐ Was war neu	- Was war neu an diesem Feedback, das Du so nicht kennst?
3.3	Was ist besser / schlechter?	☐ Evaluation	
4	**Abschluss** *Nun sind wir bereits fast am Ende des Interviews angekommen. Gerne möchte ich Dir noch ein paar Fragen zu diesem Interview stellen.*		
4.1	Wie fandst Du es, die Lehrpersonen beurteilen zu dürfen?	☐ Wohlsein / Unwohlsein ☐ Gefallen Rollenumkehrung	- Wie wohl / unwohl war Dir dabei, eine Lehrperson zu beurteilen? - Was hat Dir an dieser Rollenumkehrung gefallen? - Was hat Dir weniger daran gefallen?
4.2	Was fandst Du einfach beim Beurteilen der Lehrperson?	☐ Können ☐ Selbstvertrauen	
4.3	Was fandst Du schwierig beim Beurteilen der Lehrperson?	☐ Können ☐ Selbstvertrauen	
4.4	Wie einfach oder schwierig fandst Du die Interviewfragen?	☐ Verständnis ☐ Schwierigkeit / Komplexität	
4.5	Welche Interviewfrage war für dich besonders schwierig zu beantworten?		

Table 46 : Interview guide

F
Sample Interview Transcript

Erhebung Schüler*innenperspektive: Haupterhebung

OZ Buechenwald, 27.10.2020, Transkript B1

I: Auso. Jetz sötts loufe. Die erschti Antwort das si eifach Oudioufnahme wo du ghörsch und jetzt wechsle ich auch auf Hochdeutsch. 00:00:11

B1: Ja. 00:00:11

I: Und so, haben wir hier Aufgabenbeispiele ich/ ich spiel jetzt das erste Audio, die erste Audiodatei ab. 00:00:18

B1: Okay tiptop. 00:00:18

9039: Well, first of all I'd like to say thank you Nathalie for your presentation. I really liked it, you spoke fluently and you really explained your points. Think that was very good. First, I'd like to talk about the correctness of your speech, and I think you did a great job, you made only few mistakes and in general it was really good. It was easily understandable and the only two mistakes I heard were when you said "in place like the Vatican", it should have been "in PLACES… like the Vatican", and "wear clothes over your shoulders", I think there you went from the German sentence. Maybe you could have said, ehm "you have to HIDE your shoulders" or something like that. But in general, it was REALLY good. Then, to the content. Ehm, you chose a good example with the churches ehm in general and the Vatican specifically. And, you named your example and afterwards you explained the different points you have to avoid like, showing your shoulders, making selfies, taking pictures and other examples. I think that was very good. But, maybe you could have also, ehm, mentioned other examples. There are other places or situations where you have to be careful not just ehm visiting churches. So, and we've been talking about that quite a lot in the, inn/ last classes, so probably you remember any other situations. But… in general, you did a good job and you

explained your point well. So, I'm really happy with your presentation. Thank you very much. 00:02:20

I: So, wir können's auch jederzeit nochmal abspielen wenn du möchtest. Ich frag jetzt einfach mal mit der erst/ ich leg mal los mit der ersten Frage. Was ist dein ERSTER Eindruck vom Feedback von dieser Lehrperson? Allgemein. 00:02:36

B1: Aso, ich hab's sehr gut gefunden. Ehm, wie sie auch geredet hat also es kam sehr überzeugend vor und selbstbewusst und, dass sie halt Erfahrung mit, mit Englisch hatte. Dass sie's halt… kann. Man hat ihr's auch angemerkt. 00:02:52

I: Mhm. Und ehm, kannst du… eigentlich hast du's schon begründet, ja. Ehm, kannst du… es war ein bisschen lange aber was kannst du nochmal… was hat sie gesagt woran du dich erinnern kannst? 00:03:07

B1: Aso ich hab'… Grossteil nur gehört WIE sie's ausgesprochen hat, ehm es kam sehr viel über «general» vor und so Zeugs. Ehm aber, halt die Aussprache war sehr britisch, kann man sagen, und… ist verständlich. 00:03:27

I: Mhm. Sehr gut. Ehm, gibt es grad, gab es Teile wo du das Gefühl hattest das hab' ich nicht so gut verstanden. 00:03:35

B1: Ehm… nein, eigentlich nicht. Ehm, es liegt einfach… vielleicht wegen meinen Englischkenntnissen, aber sonst, ich denke also, sie hat GUT und deutlich gesagt. 00:03:48

I: Mhm, mhm, sehr gut. Ehm… was ehm… gibt es etwas wo du FINDEST dass es die Leh/ die Lehrperson WENIGER gut gemacht hat. 00:04:00

B1: Ehm, eigentlich nicht. Es war… SEHR gut. Es ist einfach, so, auf Dauer halt die g/… gleiche Schleife gewesen. Ehm, sonst war alles… in Ordnung. Aso dass man halt beim, wenn man eine Lehrperson ist und den Kindern was beibringen will dass man wie… d/… die Stimmung so zeigt. Also nicht einfach alles so durchredet, dass man auch mit denen sprecht… zusammen spielen kann sozusagen. 00:04:32

I: Aso so ein bisschen ein… Gespräch… 00:04:34

B1: Genau 00:04:34

I: Ja. Dass es nicht so monoton vielleicht ist? 00:04:36

B1: Das, genau. 00:04:38

I: Ja, okay. Gut, ich glaube ich verstehe was du meinst. Ehm, wenn du jetzt einfach an die Sprache von der Lehrperson denkst, also nicht unbedingt auf den TONFALL sondern einfach auf die Englischkenntnisse, ehm, dich beziehst. Gibt es/ was hat dir an der Sprache GUT gefallen. Oder weniger gut gefallen. 00:04:59

B1: Weniger gut hat ma glaub hier nichts gefallen. Aber sonst… was mir gefallen hat ist… ehm… es kam halt überzeugend vor, dass sie sich MÜHE gemacht hat. Das… hat mich so überzeugt aso man merkt dass… da Mühe drin steckt. 00:05:20

I: Mhm. Dassd ein spannender Punkt weil dann hat man das Gefühl… man ist… WICHTIG, oder? 00:05:26

B1: Ja genau, genau 00:05:27

I: Oder was man gemacht hat, ja. 00:05:27 Ehm, wie GUT denkst du, DEINER Meinung nach, spricht diese Lehrperson Englisch. 00:05:37

B1: Aso von eins bis zehn? 00:05:38

I: Mhm 00:05:38

B1: Ich denke eine… eine gute acht. 00:05:42

I: Mhm. Eine gute acht, wieso nicht eine zehn? 00:05:45

B1: Es fehlt… ich, ich glaube es fehlt… e/ etwas was für mich so eine ZEHN ist. So die Überzeugung und alles. Sie hat mich gut überzeugt, aber das mit dem Tonklang und alles, wie als hätte man das nur gelernt. 00:06:04

I: Ja, ah, ja. 00:06:05

B1: Aber, ich weiss es nicht, ich denke NOCH nich eine zehn ist es nicht, aber eine gute acht, ja. 00:06:11

I: Ja, eine gute acht. Sehr gut. Ehm. Ich, ich spiel noch mal die ersten so zwanzig Sekunden oder so ab, einfach damit wir's noch ein bisschen klarer nochmal vor Augen haben, weil ich stell ein paar präzisere Fragen jetzt noch. 00:06:27

9039: Well, first of all I'd like to say thank you Nathalie for your presentation. I really liked it, you spoke fluently and you really explained your points. Think that was very good. First, I'd like to talk about the correctness of your speech, and I think you did a great job, you made only few mistakes and in general it was really good. It was easily understandable and the only two mistakes I heard were when you said "in place like the Vatican", it should have been "in PLACES… like the Vatican", and "wear clothes over your shoulders", I think there you went from the German sentence. Maybe you could have said, ehm "you have to HIDE your shoulders" or something like that. 00:07:18

I: So, ehm, nochmal kurz zur Sprache. Ehm, woran erkennst du, denkst du, dass diese Person GUT im Englisch ist? 00:07:31

B1: Ehm, also a/ an der Aus/ AUSsprache, merkt man das, und, wenn sie, halt, eine überzeugende Stimme dazu hat. Dass… daran merkt ma/ man, dass, s/ s/hat dass sie's kann. 00:07:49

I: Ja, also hast du das Gefühl es hat auch viel mit dem Tonfall zu tun 00:07:54

B1: Aso mit dem Tonfall, wie sie auch redet, wie sie das rüberbringt… ja. 00:07:59

I: Mhm, mhm, okay. Cool. Wa/ wie war wie anstrengend war es für dich, die Lehrperson zu verstehen? 00:08:06

B1: Im ersten Part hab' ich nur auf… die Tonlage und halt WIE sie redet, geachtet. Aber im zweiten, im zweiten Durchlauf hat sie die kurzen dreissig Sekunden, hab' ich auch genau auf die Worte geachtet und ich konnte sie gut verstehen, und hab… fast vieles/ aso fast alles mitbekommen. 00:08:32

I: Mhm. Super. Ehm. Ich geh' weiter. Du hast es schon mal ANgesprochen, ich geh' jetzt noch ein pa/ ein bisschen spezifischer darauf ein – wie hat die AUSSPRACHE von Eng/ von der englischen Sprache für DICH geklungen? 00:08:48

B1: Die Aussprache von… 00:08:51

I: Von dieser Person. 00:08:52

B1: … von dieser Person. 00:08:52

I: Genau 00:08:53

B1: Ehm, sehr… sehr britisch, wie gesagt. Halt, ehm, als… ich denke sogar dass sie irgendwie vielleicht sogar eine Englische Familie hat oder vielleicht aus dem England kommt oder, ehm… w/ wie sagt man das, wenn sie dort in die Schule kurz war 00:09:14

I: Mhm einen Austausch vielleicht? 00:09:14

B1: Ein Austausch, so etwas, dass sie halt mal was… ausser SCHULISCH halt, mal was mit Englisch zu tun hatte, so. 00:09:24

I: Mhm. Aso den Akzent, so wie sie klingt, klingt irgendwie… 00:09:28

B1: Genau, als kann sie so SWITCHEN. 00:09:30

I: Ja. Findest du das… w/ was hältst du davon? Findest du das GUT, oder 00:09:34

B1: Ich finde das GUT, dass man halt, wenn man eine Sprache… unterrichten will, dass man auch so zeigt wie ist es wirklich ist. Was bringt es mir wenn ich in der, wenn ich die Deutsch, das Deutsch-Englisch in England anwende. Ist da halt... 00:09:51

I: Mhm, mhm. Ja. Guter Punkt. Ehm, was denkst du, wie GUT würde eine Person DIESE Lehrperson verstehen die jetzt nicht so gut im Englisch ist aso in der Schule. 00:10:03

B1: Ehm 00:10:06

I: Jemand… der MÜHE hat. 00:10:06

B1: Ich denke eine vier, vier von zehn, fünf von zehn, so. 00:10:16

I: Mhm. Weshalb, woran denkst du könnte es LIEGEN, oder was würde die Person schwierig finden… 00:10:23

B1: Ehm, aso diese Person hat sehr… hoche Wörter, und nicht die einfachen Wörter, die jeder halt, das Basic, das hat sie halt nicht angewendet sondern das, was man eigentlich gelernt hat, lernen muss oder lernen muss. Aso ich hab' da nicht so basic Wörter gehört sondern. Da muss man schon ein bisschen lernen dass man so Wörter kann. 00:10:46

I: Ja, aso denkst du, es w/ WIE hilfreich wäre es oder inwiefern denkst du wäre es hilfreich wenn die Person ihr Englisch dem Niveau von dem Schüler, oder der Schülerin, anpassen würde? 00:10:58

B1: Wie gut sie wäre? 00:10:59

I: Ja 00:10:59

B1: Ich denke eine neun. 00:11:01

I: Ja, eine n/ okay. Sehr gut. 00:11:02

B1: So 00:11:03

I: Aso könnten wir vielleicht jetzt sagen, würdest du denken, das NIVEAU jetzt von dieser Person die wir gerade gehört haben, wie sie spricht ist eher… ehm… KOMPLEX, ein bisschen, du hast gesagt ho/ eine ho/ hohe Sprache so 00:11:18

B1: Genau, so, eine… wie kann ma/ das ist eigentlich das BASIC halt. Es ist mehr gebildet, so. Das gebildete Englisch, so. Es gibt das basic Englisch und das gebildete Englisch und ich kann nur zum Beispiel das Basic und dazwischen sind zwei. 00:11:35

I: Mhm, mhm, ja, sehr gut. Ehm, das geht auch grad n' bisschen in meine nächste Frage rein und zwar in den WORTschatz von der Lehrperson die wir gehört haben. Ehm, findest du, dass die Person die Wörter PASSEND gewählt hat, dass die TREFFEND waren, oder war da, waren da Dinge drin die du denkst das wa/ das war FALSCH. 00:11:57

B1: Nein ich hab eigentlich ALLES, oder fast alles, GUT verstanden und, ehm, so würd ich auch die Sätze sagen, aso, rüberbringen. 00:12:07

I: Mhm, aso es war nicht so ds/ dass du jetzt grad gedacht hast dass da das war jetzt das hat sie falsch gemacht oder so 00:12:13

B1: Nein 00:12:13

I: Nein? GUT. Ehm, jetzt zur, äh, FLÜSSIGKEIT, wie sie gesprochen hat. Wie FLÜSSIG fandst du das Englisch dieser Lehrperson? 00:12:24

B1: Ehm… es war… MEHR als mittelmässig, es war nicht SEHR flüssig, und auch nicht GAR NICHT flüssig, es war MEHR als mittelmässig, aso es… FEHLT noch ein bisschen ah si/ wa/ GANZ ein bisschen unsicher, aber… ja. 00:12:47

I: Ja. Gut. Ehm und das Tempo, wie fandest du das? 00:12:50

B1: Find ich… in Ordnung, also... 00:12:51

I: War gut. 00:12:54

B1: … perfekt so. 00:12:54

I: Ja. Sehr gut. 00:12:55

B1: Nicht zu schnell, nicht zu langsam, man bekommt es mit über. 00:12:57

I: Gerade so… richtig. 00:12:59

B1: Genau 00:12:59

I: Wunderbar. Ehm, gibt es etwas wo du denkst die Lehrperson… müsste an DEM noch arbeiten, damit sie noch BESSER wird im Englisch. 00:13:10

B1: Ehm… diese Person soll einfach… SICHERER sein mit sich selber. Dass sie… ehm… einfach… das redet als wär das… ihre Muttersprache. Also aber sonst find ich alles gut daran. 00:13:28

I: War alles gut. Super. Ehm. Möchtest du sonst noch etwas sagen zu dem was du jetzt gehört hast? Zu dieser Aufnahme? 00:13:38

B1: Aso, ich find's einfach so in Ordnung, war gut und… 00:13:41

I: War in Ordnung so? Sehr gut. Ehm, dann… noch ganz kurz… Im ENGLISCHunterricht, in welcher Sprache gibt dir DEINE Lehrperson Rückmeldung? 00:13:54

B1: Zu mir in Englisch. 00:13:56

I: In Englisch, ja? Und ist ds, war das jetzt GROSS anders, das was du jetzt gehört hast zu der Rückmeldung die du dir gewohnt bist? 00:14:07

B1: Mh… wie kann ich das sagen, es ist… es war SCHON anders, definitiv, aber nicht dass ich nicht das nie gehört habe oder NIE verstehen, nie verstehen könnte. 00:14:23

I: Ja. 00:14:24

B1: Aber ja, es hat schon ei/… Unterschied. 00:14:27

I: Ja, ja? Kannst dus sogar sagen was anders ist oder ist's einfach ein Gefühl? 00:14:30

B1: Ehm, das ist das Gefühl und, ehm, zwei drei Wörter die man halt nicht… ja 00:14:37

I: … die ein bisschen… unbekannt… 00:14:40

B1: Genau, aber sonst habe ich alles mitbekommen. 00:14:40

I: Okay, alles klar. Gut, dann, äh, machen wir noch einmal eine Aufnahme, ist das in Ordnung? 00:14:50

B1: Tiptop 00:14:50

I: Gut, und zwar… gehen wir da rein… 00:14:56

8842: Thank you, Nathalie, for your short presentation. So, you did that well, you used examples like the churches and speciflically the Vatical, and so you had a huge content so for example need to wear clothes over the shoulers and that you're not allowed to be, to take photos. Ehm… we h/… and you had ah used… a wide range of vocabulary. We had to look dear ehm sometimes of the pronunciation like "the" and "the" and some verb forms like "take photos" instead of "make photos" but… you… have spoke very good… well. 00:16:03

I: So, das war jetzt jemand ANDERS. Was ist DEIN erster Eindruck von DIESER Aufnahme? 00:16:13

B1: Aso, ehm, die Aufnahme ist nicht gut. So. Ehm, sie hat SEHR lange gebraucht bis sie die Wörter gefunden hat. Und auch… sehr gezögert, nicht fliessend… und halt auch die T/, der Ton/ die Tonlage halt, das war nicht… 00:16:38

I: Mhm, hat dich nicht überzeugt 00:16:38

B1: Nein, aso 00:16:41

I: Ja, ehm. Kannst du, gab es Dinge die du verstanden hast oder eben NICHT verstanden hast? 00:16:47

B1: Es war alles n' bisschen… aso, dreingeredet und… LÄRM, und ma/ man kann das nicht mit, aso, s/ sist war nicht wie das beim Ersten, das ist nur eine Person geredet hat, klar und flüssig und genau die Lautstärke, da kamen bei ihr «ehms» rein und «meeh» und… da bekommt man halt nicht alles mit über. Wenn jemand flüssend redet bekommst du WORT FÜR WORT den ganzen Satz mit. Ja halt, bis sie fertig ist dann musst du das andere verarbeiten, mit… 00:17:24

I: Mhm, aso, wüsstest du jetzt was du machen müsstest nachdem sie… das gesagt hast oder/ hat oder wärst du ein bisschen… was mach' ich jetzt mit dem ich hab's nicht verstanden. 00:17:34

B1: Aso ich hab's nicht verstanden. Es hat kein Sinn gemacht. 00:17:39

I: Mhm. Ehm… Gibt es etwas, was die Person GUT gemacht hat? In deinen Augen? 00:17:48

B1: Aso, ich denke ihr Englisch ist… GUT. Ehm, ich weiss nicht wieso sie erstens so neu/ so nahe an dem Mikrofon geredet hat, und, ehm wieso sie so lange gebraucht hat das rüberzubringen, aber sonst… war halt der Rest nicht so meins. 00:18:08

I: Mhm. Okay. Ehm. Du hast schon gesagt, eh, was sie weniger gut gemacht hat so die vielen «ehms», das Zögern, die Lautstärke und so, gibt es sonst noch etwas was du findest das hat sie nicht gut gemacht? 00:18:21

B1: Mh es waren zu viele Pausen. Dazwischen. 00:18:25

I: Mhm. Aso ein bisschen schwierig zum, damit man dem überhaupt FOLGEN kann weil es so viele… 00:18:30

B1: Genau 00:18:30

I: Alles klar. Ehm, wenn du jetzt, NUR an die, das ENGLISCH von dieser Lehrperson wieder denkst also nur an den Sprachgebrauch, gibt es, was hat dir an der Sprache GUT, oder eben NICHT gut, gefallen? 00:18:45

B1: Es war ein… angenehmer Akzent. Aber… w/ wie sie's rübergebracht hat war nicht in Ordnung für mich aso was, ehm mit dem ZÖGERN und alles… drum und dran 00:19:00

I: Und dann hilft der schöne Akzent halt auch nicht so viel wenn… 00:19:04

B1: Genau 00:19:04

I: Ja, alles klar. Ehm, du hast scho/ vorher AUCH schon angesprochen, du denkst das Englisch ist GUT von dieser Lehrperson, wie GUT denkst du is/ ist sie, so jetzt vielleicht nochmal… 00:19:15

B1: eine… 00:19:15

I: …auf der eins bis zehn 00:19:16

B1: sechsein/ sechseinhalb, sieben, so. 00:19:18

I: Mhm. Ehm, woran erkennst du, dass sie noch nicht so gut ist, denkst du? 00:19:24

B1: Sie hat SEHR lange gebraucht bis sie die passenden Wörter gefunden hat. Und… sie war halt, stockern, also sie hat gestockert und hat viele Pausen gemacht. Da ist noch nicht eine zehn da. 00:19:42

I: Ja. Kann ich gut verstehen (lacht). Ehm, fandest du es anstrengend, zuzuhören? 00:19:47

B1: Ja es man muss sich schon konzentrieren dass man was mitbekommt. 00:19:52

I: Ja. Ääh, gut, dann die… Aussprache haben wir schon abgedeckt… Eine Person, die nicht so gut im Englisch ist, wie gut würde die Person diese Lehrperson verstehen, denkst du? 00:20:06

B1: Mh ja ich denke sie würde eine sieben also. Weil die auch sehr langsam redet und, ehm, man bekommt es mit. Denk ich. 00:20:25

I: Mhm, aso würde, würde eigentlich diese Person TROTZDEM gut verstehen können 00:20:30

B1: Verstehen SCHON denk ich, genau für so etwas ist das gut denk ich mal. 00:20:35

I: Ja, aso jemand der NICHT so gut Englisch kann… 00:20:38

B1: Weil dass sie mit ihm langsam reden kann und da braucht sie ihre Zeit bis sie alles versteht dann kann sie auch ihre Zeit nehmen, ja. 00:20:44

I: Ahhh, ja. Aso ein bisschen BESSER als die Person vorher? 00:20:47

B1: Genau. Aso… wie kann ich sagen… nein, ich denke nicht dass die Person jetzt besser ist als die andere, was sie, wie sie's den, einem… Kind beibringen kann das/ der… nicht so gut Englisch kann. Aber, sie könnte es AUCH gut, denk ich mal. 00:21:08

I: Ja, alles klar. Aso sie wüsste wahrscheinlich wie… 00:21:11

B1: Genau, wie, genau. 00:21:11

I: Mhm. Ehm, wenn du an die WÖRTER, denkst, die die Person, die jetzt diese Lehrperson benutzt hat, ehm, waren die, ist dir da etwas aufgefallen beim WORTschatz dass etwas NICHT passend war oder grundsätzlich… 00:21:25

B1: Nein, es war schon alles in Ordnung, aber, ehm, es war halt, sehr LEISE und halt nicht… so fliessend halt gewonnen es war… 00:21:37

I: Mhm, mja. Da beantwortest du auch schon meine nächste Frage, eben dass das hast du schon erwähnt, sie hat geSTOCKT und es war nicht so FLÜSSIG wie, jetzt im Vergleich zur vorherigen. 00:21:48

B1: Genau 00:21:48

I: Mhm. Wie fandest du das Tempo? 00:21:50

B1: Das Tempo war langsam. Nicht ZU langsam aber ein bisschen langsamer als das in als das perfekte Mitte. 00:21:58

I: Ja. Also für dich persönlich wär es ZU langsam? 00:22:01

B1: Ein bisschen langsam, ja wahrscheinlich. 00:22:04

I: Würdest du dich langweilen 00:22:06

B1: Aso für mich, ich hätt es gern mittelmässig. Nicht zu schnell nicht zu langsam. 00:22:08

I: Ja. Ja. Ehm, woran müsste diese Lehrperson arbeiten, um BESSER im Englisch zu werden? 00:22:16

B1: Aso da braucht sie auch Selbstvertrauen, dass sie, und ein aufgewecktes ICH dass sie halt, das wie ein… dass sie den Rhythmus hat, wie sie reden soll und… ja. 00:22:30

I: Mhm, mhm. Ehm, gibt es sonst noch etwas, gibt es einen TIPP, oder so die du dieser Person geben würdest? 00:22:37

B1: Mh nein, mit fällt nichts ein. 00:22:38

I: Fällt nichts ein, ja. Gut. Äh, sonst noch etwas zu dieser Aufnahme, so? 00:22:44

B1: Nein, nicht 00:22:44

I: Nein. GUT. Dann hätten wir eigentlich beide Aufnahmen schon besprochen. Dann würd' ich einfach noch ganz kurz zur, zum Interview allgemein, ehm, ein paar Ffragen stellen. Wie war es für dich, Lehrpersonen beurteilen zu dürfen? 00:23:00

B1: Es war… in Ordnung halt. Vielleicht helf' ich damit weiter, vielleicht auch nicht, ja 00:23:10

I: Ja 00:23:10

B1: Ehm, ist auch nicht BÖSE gemeint wenn ich was… angriff/ angriffliches gesagt habe. Ist nur MEINE Meinung und… ja, und ich hoffe mit dem was ich gesagt habe dass sie ehm daraus lernen können, also dass sie daraus, wi/ dass sie verstehen, wie es rüberkommt so einer… Schüler, Schülerin, so 00:23:35

I: Das auf jeden Fall also das war… SEHR SEHR hilfreich, und, vor allem auch für mich sehr, sehr spannend oder. Mich interessiert das natürlich SEHR. 00:23:43

B1: Ja klar, richtig 00:23:44

I: (lacht) ja. Das ist eh, und auch, aso ich fand nicht dass da irgendetwas Angriffliches dabei war, und eh, eben GENAU deine Meinung ist eben gerade so wichtig, aso ich find' das SUPER. Ehm, fandst du es, wie/ we/ fandst du es einfach oder schwierig Lehrpersonen zu beurteilen nach ihrer, nach ihrem Englisch-Können, Wissen? 00:24:06

B1: Ehm, aso es war EINFACH, zuzuhören, aber ehm, wie zu BEURTEILEN, dass es… HÖFLICH klingt und halt, dass es, dass ehm, wie… KLAR klingt dass ich wie es ihnen wie ich es meinen soll, DAS war ein bisschen schwieriger, würd' ich sagen. 00:24:31

I: Ja. Und zu erkennen, wie gut dass die im Englisch sind, wie war das für dich? 00:24:35

B1: Eben, das ist halt auch die andere Sache wo, ehm, nich, für mich SEHR schwierig ist, weil ich halt noch ein Schüler bin und noch nicht so hochgebildet bin wie SIE im Englisch. Ehm, da kann ich halt nicht sehr viel sagen dass das einfach für mich ist. So. Es braucht schon… die Zeit um das zu Verstehen mitzubekommen und dann kann ich erst die Rückgabe geben. Oder Rückmeldung, ja. 00:25:03

I: Absolut, kann ich sehr gut verstehen. Ja. Sehr gut. Ehm, gab es noch zum/, gab es Interview Fragen die du besonders schwierig fandst oder wie war das jetzt ALLGEMEIN so in diesem Gespräch für dich? 00:25:13

B1: Das GESPRÄCH war sehr angenehm und ehm, hab wieder dazu was gelernt, ehm was ich SCHWER fand ist ehm die Fragen halt wenn, wie ich diese Person finde zum Beispiel was sie richtig gemacht hat was sie falsch gemacht hat, das noch aufzuteilen, und dann, ja 00:25:39

I: So die einzelnen Details, so 00:25:41

B1: Genau, die genauen Details eben darüber und so Sachen. 00:25:44

I: Ja, das ist auch eine schwierige Aufgabe. Also ich, ich hab' dich schon ein bisschen gefordert (lacht). GUT. Hey, vielen, vielen Dank 00:25:52

B1: Immer gern 00:25:53

I: Ich stell jetzt mal die, das Aufnahmegerät, AUS. 00:25:56

B1: Ja 00:25:5

G Coding Frame Sub-Study

Coding Conventions

This section presents the coding frame developed and used for the qualitative sub-study. It includes the following instructions (cited from the original coding frame):

- Double coding is possible and permitted, for example in cases where there is an (obvious) semantic overlap (e.g., ambiguity of meaning). The following extract represents such an example. This turn was both coded as "RQ1-24: ÜS Überzeugende Stimme und Ausdruck" and "RQ1-7: SKNS Native-Speakerism":

  Example 1: B1: Es fehlt… ich, ich glaube es fehlt… e/ etwas was für mich so eine ZEHN ist. So die Überzeugung und alles. Sie hat mich gut überzeugt, aber das mit dem Tonklang und alles, wie als hätte man das nur gelernt. 00:06:04

- The smallest coding unit constitutes a semantic unit. A semantic unit in this context is defined as a distinct meaning component comprised of linguistic signs. In other words, a semantic unit in the present context represents a clear meaning component in the text or a unit of meaning that is in itself coherent and self-contained; i.e. if decoupled from its context, it is still understandable in terms of its meaning. Thus, a semantic unit may encompass anything ranging from a phrase to several turns between the interviewer and the interviewee:

  Example 1: B1: es war halt, sehr LEISE

  Example 2: B1: Das Tempo war langsam. Nicht ZU langsam aber ein bisschen langsamer als das in als das perfekte Mitte. 00:21:58

  I: Ja. Also für dich persönlich wär es ZU langsam? 00:22:01

B1: Ein bisschen langsam, ja wahrscheinlich. 00:22:04

I: Würdest du dich langweilen 00:22:06

B1: Aso für mich, ich hätt es gern mittelmässig. Nicht zu schnell nicht zu langsam. 00:22:08

- There are five interview transcripts in total based on five semi-structured guided interviews. The length of the interviews ranges from 22 to 33 minutes.
- The coding frame was developed deductively and inductively. The deductive categories were derived from the interview guide and the assessment criteria from the profession-related language competence assessment rubric (PRLC-R). The inductive categories were derived from the textual material itself.

Coding Frame

Abbreviation	Thematic main category	Subcategory	Definition	Examples
Feedback als sprachliche Produktion				
SK	Sprachkompetenz			
RQ1-1: SKAllg		Sprachkompetenz: Allgemein	Wahrnehmung der Sprachkompetenz allgemein: Aussagen zum Gesamteindruck oder Ersteindruck, Aussagen zur allgemeinen Sprachkenntnis der LP	Aso, ich hab's sehr gut gefunden. Ehm, wie sie auch geredet hat also es kam sehr überzeugend vor und selbstbewusst und, dass sie halt Erf Erfahrung mit, mit Englisch hatte. Dass sie's halt… kann. Man hat ihr's auch angemerkt. an der Aus/ AUSsprache, merkt man das, und, wenn sie, halt, eine überzeugende Stimme dazu hat. Dass… daran merkt ma/ man, dass, s/ s/hat dass sie's kann.

				I: wie GUT denkst du, DEINER Meinung nach, spricht diese Lehrperson Englisch. 00:05:37 B1: Aso von eins bis zehn? 00:05:38 I: Mhm 00:05:38 B1: Ich denke eine… eine gute acht. 00:05:42
RQ1-2: SKERF		Erfahrung	Erfahrung mit der Sprache, sprachliche und kulturelle Versiertheit basierend auf erfahrungsbedingten Faktoren in der Sprachbiographie der LP wie Auslandsaufenthalte, ausserschulische/ extraprofessionelle Aktivitäten im Zusammenhang mit der Sprache	dass sie halt Erfahrung mit, mit Englisch hatte. Dass sie's halt… kann. Man hat ihr's auch angemerkt Ein Austausch, so etwas, dass sie halt mal was… ausser SCHULISCH halt , mal was mit Englisch zu tun hatte, so.
RQ1-3: SKKOR		Korrektheit	Sich sprachlich korrekt ausdrücken: Aussagen zu Grammatik (Syntax, Morphologie)	I dengg dasch, da ghöri gad vo usch vo i weiss nöd, jo. Wenn öpper so schnell Englisch redt oder so, jo halt eifach so, halt so richtig gueti Uusproch het und n/ kein einzige Fehler macht, DENN
RQ1-4: SKWS		Wortschatz	Aussagen zum Wortschatz und Wortwahl im Allgemeinen (im Gegensatz zu «adressatengerechtem Wortschatz»), sich im gegebenen Kontext mit inhaltlich passender Wortwahl ausdrücken, Aussagen	i weiss nöd öb& si vilicht paar Fehler gmacht het, aso i weiss nöd, i sege immer im Englische «pictures», und si brucht «photos» gseit, aber, s Meitli im Video het au gseit ehm «selfies and photos», aber i glaub da isch en Fehler v/ vom Meitli gsi, wil si immer gseit het «pictures, selfies and pictures».

			zu Wortschatztiefe und Wortschatzbreite allgemein	Und d Lehrperson het au gseit «photos» und ez, jo. Aso ich denke sie hat vielmal das Wort «also» oder, halt, die gleichen Wörter einglich benutzt.
RQ1-5: SKFL		Flüssigkeit	Aussagen zur Sprechflüssigkeit, Zögern und Sprechpausen, sowie zum Einsatz von Füllwörtern und weiteren Strategien zur Pausen-überbrückung.	es war… MEHR als mittelmässig , es war nicht SEHR flüssig, und auch nicht GAR NICHT flüssig, es war MEHR als mittelmässig, aso es… FEHLT noch ein bisschen ah si/ wa/ GANZ ein bisschen unsicher
RQ1-6: SKSG		Sprechgeschwindigkeit	Sprechgeschwindigkeit, Aussagen dazu, ob und wie sich eine LP in angemessenem Tempo ausdrücken kann, Aussagen zur persönlichen Präferenz der Sprechgeschwindigkeit der LP, Aussagen zum Rhythmus der Sprachproduktion über syntaktische Einheiten hinweg	Aso ich fand es dass es ehm angenehm war. Auch dass s/ dass man auch die Wörter noch hör/ aso hören konnte sozusagen, ja. Dass sie den Rhythmus hat, wie sie reden soll und… ja.
RQ1-7: SKNS		Native-Speakerism	Aussagen die implizieren, dass ein Muttersprachlerniveau (nicht) anstrebenswert ist und mit hoher Sprachkompetenz (=proficiency) gleichgesetzt wird; wenn eine LP ein von den SUS wahrgenommenes Muttersprachlerniveau spricht und die SuS diese LP als	Ich finde das GUT, dass man halt, wenn man eine Sprache… unterrichten will, dass man auch so zeigt wie ist es wirklich ist. Was bringt es mir wenn ich in der, wenn ich die Deutsch, das Deutsch-Englisch in England anwende .

			kompetenter einschätzen als jene mit einem deutschsprachigen Akzent (die aber genauso sprachlich kompetent sind); Annahmen, dass man nur auf Muttersprachlerniveau komplett «richtig» sprechen und komplett sprachkompetent sein kann.	
Feedback aus Sicht der Rezeption: Feedback als Interpretation				
VER	Verständnis			
RQ1-13: VERAllg		Verständnis: Allgemein	Aussagen dazu, zu welchem Grad die SuS die LP allgemein verstanden haben; allgemeiner Gesamteindruck; inhaltliche Wiedergabe des FB durch die SuS zum Abschätzen bzw. Überprüfen, was und wie viel die SuS verstanden haben.	im zweiten Durchlauf hat sie die kurzen dreissig Sekunden, hab' ich auch genau auf die Worte geachtet und ich konnte sie gut verstehen, und hab… fast vieles/ aso fast alles mitbekommen. I: gab es Teile wo du das Gefühl hattest das hab' ich nicht so gut verstanden. 00:03:35 B1: Ehm… nein, eigentlich nicht. Ehm, es liegt einfach… vielleicht wegen meinen Englischkenntnissen, aber sonst, ich denke also, sie hat GUT und deutlich gesagt.
RQ1-14: VERANS		Anstrengung	Aussagen dazu, ob und wie sehr sich die SuS anstrengen oder konzentrieren mussten, um die LP zu verstehen. Dazu gehört «gut	Ja es man muss sich schon konzentrieren dass man was mitbekommt.

			hinhören müssen», «gut aufpassen müssen» etc.	
RQ1-15: VERSL		Verständlichkeit schwache Lernende	Aussagen zur Einschätzung der SuS, wie verständlich die LP für schwache Lernende wäre	Aso wenn sie's, aso ich glaube schon dass man ein Teil verstehen kann. Auch wenn man nicht SO gut Englisch kann. Ehm dass man's… ehm, sich vorstellen kann was sie sagt und… ja.
AUS	Verständliche Aussprache			
RQ1-8: AUSAllg		Verständliche Aussprache: Allgemein	Sich mit verständlicher Aussprache und Betonung ausdrücken, Aussagen zur Verständlichkeit und zur (un)klaren, (un)deutlichen, (un)zugänglichen Aussprache	Aso, sie hat… eine gute Aussprache im Englisch. Ehm, sehr deutlich, und auch… aso m/ auch wenn sie ein bisschen schnellerer sprechen würde sie auch DANN noch verstehen. Aso ich hab'… Grossteil nur gehört WIE sie's ausgesprochen hat, ehm es kam sehr viel über «general» vor und so Zeugs. Ehm aber, halt die Aussprache war sehr britisch, kann man sagen, und… ist verständlich.
RQ1-9: AUSVAR		Varietät	Aussagen zu subjektiven Zuordnungen zu Sprachvarietäten, hier definiert als das Übertragen / Übernehmen von Aussprachegewohnheiten aus einer anderen Sprache (bspw. der Mutter-/Erstsprache) in Englisch, sprachhintergründliche Färbung	die Aussprache war sehr britisch, kann man sagen, und… ist verständlich. Abo s isch halt da, da, da Englisch wo me nöd wött zuelose so da, wenn me döt ine wie die Schuel wür hogge und da wür lose wür me so dengge… so…

			der Aussprache in Englisch (z.B., im Englisch eine Schweizer, oder aber auch eine Amerikanische oder Britische Färbung haben); Auch Aussagen zum Mögen / Nicht-Mögen bestimmter Sprachvarietäten	
RQ1-23: SB	Selbstbewusstsein		Allgemeines und sprachlich selbstbewusstes und selbstsicheres Auftreten, Aussagen zur subjektiv wahrgenommenen Selbstsicherheit / Selbstvertrauen / Selbstbewusstsein	Ehm… diese Person soll einfach… SICHERER sein mit sich selber. Aber… w/ wie sie's rübergebracht hat war nicht in Ordnung für mich aso was, ehm mit dem ZÖGERN und alles… drum und dran
ASW	Auditive Sprachwahrnehmung: Psychophonetik / Psychoakustik			
RQ1-10: ASWTF		Sprechmelodie / Tonfall als parasprachliche Funktion der Prosodie (Ebene der A-Prosodie)	Subjektiv wahrgenommener Klang der Stimme, der einen Gemütszustand/eine Laune, eine Stimmung oder eine Eigenschaft der LP wiedergibt und entsprechend Gefühle als Reaktionen bei den SuS triggert (z.B. Sympathie, Apathie)	Also sie tönt sehr freundlich nächste Mal sollte sie ein bisschen glücklicher sein und nicht so müde
RQ1-11: ASWTH		Subjektiv wahrgenommene Tonhöhe (pitch): Mel	Subjektiv wahrgenommene Intonation im Sinne der Stimmführung und des Tonhöhenverlaufs über	Aber i han s Gfühl es isch so e, es Flüschtere, denn no ingwie so e, so chli so e wie e so e Stöhne, so, so e, so am

			sytnatktische Einheiten (= Tonheit)	Schluss ischs immer so chli so UFE gange.
RQ1-12: ASWLH		Subjektiv wahrgenommene Lautheit (loudness): Sone	Aussagen über die subjektive Wahrnehmung und Empfindung der Lautstärke (= Lautheit), in der die LP gesprochen hat.	es war halt, sehr LEISE sisch au gnueg luut gsi. also ich finde dass sie hat auch ein bisschen… LEISE gesprochen weil ich musste ein bisschen zu nah kommen zu dem Computer, weil ich habe es nicht so gut gehört und verstanden dann, also
Feedback als pädagogisches Werkzeug				
PK	Pädagogisches Wissen (PK)			
RQ1-28: PKFB		Pädagogisches Wissen: FB	Aussagen dazu, wie die LP ihre Kenntnisse über die Feedbackkultur, die sich auf die Gestaltung von Unterrichtssituationen beziehen, umsetzt, und die fachunabhängig, das heißt auf verschiedene Fächer und Bildungsbereiche anzuwenden sind	dass sie nicht gad mit dem Schlechten angefangen hat sondern mitem Guten. Ja das find ich ehm s/ angenehm wenn man das so hört von den Lehrperson nein, ich denke nicht dass die Person jetzt besser ist als die andere, was sie, wie sie's den, einem… Kind beibringen kann das/ der… nicht so gut Englisch kann. Aber, sie könnte es AUCH gut, denk ich mal. sie war… MEHR nett, also sie war nicht so KRItisch. Mh, sie sollte schon ein bisschen kritisch sein

RQ1-29: PKERKL		Erklären: Methode	Gewählte Methode zum Erklären als Subkategorie von Pädagogischem Wissen; Aussagen dazu, wie die LP "Erklären" methodisch gelöst hat und wie gut das bei den SuS angekommen ist.	Nöd mitem Verstendnis, wenns die Person LANGSAM gseit het und vilicht au ade Wandtafele aso i finds immer am Beschte, wenn MIO, wel es git ebe mega vil Lehrer wo da NÖD machet, si gönd, si sprechet eifach und si zeigets nöd ade Wandtafele. Me mues denn ade Wandtafele mit de Grammatik erkläre wür, denn würis würklech verstoh wenn döt hets au mitem spreche heti iz au nüt verstande wemmor denn aneschribt und denn nomol churz erchlärt denn gohts würklech guet. Aso i han ez ghört dass si gar nöd d Fehler erchlärt het. So, so NULL so, irgendwie so, je in sinere Sicht het si glaub fasch kai Fehler gmacht. Si het nämlich glaub nu gseit dass sis guet findet dass si got go Bilder mache. Ehm, aber, jo si het Fehler gmacht aso d Drittklässleri, und da muesch ihre au erchläre, wil, wenn sis jo nöd weiss denn macht sis jo immer falsch. Sie hatte einfache Wörter und auch immer eine Begründung dazu gesagt.

				und auch, wie sie's ja so erklärt wo die Fehler waren und so, hat sie auch gut erklärt.
AB	Adressatenbezug			
RQ1-16: ABAllg		Adressatengerechtigkeit: Allgemein	Aussagen dazu, ob und wie sich die LP den Lernenden gegenüber allgemein verständlich ausdrückt, bswp. durch das Vereinfachen der Wortwahl oder der Verlangsamung der gesprochenen Sprache; Aussagen dazu, wie sich die LP dem Kontext des Klassenzimmers und dem Niveau der SuS angemessen ausdrückt oder angepasst hat.	Weil die auch sehr langsam redet und, ehm, man bekommt es mit. Verstehen SCHON denk ich, genau für so etwas ist das gut denk ich mal.
RQ1-17: ABWS		Adressatengerechter Wortschatz	Allgemeine Aussagen dazu, ob und wie sich die LP den Lernenden gegenüber mit angepasstem Wortschatz angemessen und verständlich ausdrückt	aso diese Person hat sehr… hoche Wörter, und nicht die einfachen Wörter, die jeder halt, das Basic, das hat sie halt nicht angewendet sondern das, was man eigentlich gelernt hat, lernen muss oder lernen muss. Aso ich hab' da nicht so basic Wörter gehört sondern. Da muss man schon ein bisschen lernen dass man so Wörter kann. B4: Dass sie einglich ein Deu/ eh ein Englisch geredet hat das ich einglich auch gu/ also gut verstehe nicht irgendwie solche englische Sachen die

				ich nicht verstehe so sch/ komplizierte Wörter. I: Ja, also meinst du mit komplizierten Wörtern meinst du Wörter die du noch nicht kennst die ganz… B4: Ja oder dass zum Beispiel ein Synonym, das viel schlimmer also nicht schlimm aber unverständlicher ist das meine ich kein.
RQ1-18: ABKom		Adressatengerechte Komplexität	Aussagen dazu, ob und wie sich die LP den Lernenden gegenüber mit angepasster sprachlicher und inhaltlicher Komplexität ausdrückt (cf. CEFR-CV, Council of Europe, 2018, p. 126: «Mediation strategies to explain a new concept […]: Adapting Language […]: ► paraphrasing; ► adapting speech / delivery; ► explaining technical terminology».), Aussagen zu Register	das ist eigentlich das BASIC halt. Es ist mehr gebildet, so. Das gebildete Englisch, so. Es gibt das basic Englisch und das gebildete Englisch und ich kann nur zum Beispiel das Basic und dazwischen sind zwei. Ehm… I dengg die Person hetts würkli… fasch nöd verstande. Wells ebe mit de, mit dee… Textzemesetzig z tue het. Ja wenniz zum Bispil die Schüelerin wür froge, ehm, wa meinet sii, aso, döt wosi gseit mit da… Witz… ööh, weiss nöd weles da gsi isch mit eehm… aber halt en Fehler wo si gmacht het, und d Schüelerin wür froge wa heisst da, denn wüssti nöd öbs, öb sis weiss wel es got halt döt, da isch halt Grammatik, aber i denk scho, i cha iz jo nöd so vil

				sege wel si nur halt iz so paar, ei Minute füfzg so was gseit het und ich halt nid vil würklech weiss öb sis… no besser chan. Ehm vilicht probiere dass mor, ehm, zersch luegt dass, aso da mitem Lobe am Afang isch recht guet, aber dass mr denn döt scho i de Mitti weg de Fehler, dass mo döt nomol ganz churz chli, vilicht eifacheri Wörter nimmt oder halt eifach chli EIFACHER macht und eifach so… PROBIERT eifach erchläre.
RQ1-19: ABPort	Inhalt adressatengerecht portionieren		Aussagen dazu, ob und wie die LP komplexen Inhalt proportioniert und adressatengerecht herunterbricht (cf. CEFR-CV, Council of Europe, 2018, p. 127: «Mediation strategies to explain a new concept […]: Breaking down complicated information […]: ► breaking a process into a series of steps; ► presenting ideas or instructions as bullet points; ► presenting separately the main points in a chain of argument")	Mhm, aso ebe zum Bispil jetz irgendwie… i het zum Bispil ez irgendwie agfange gha, und het nomol segs nomol, und denn hetti si gstoppt. Oder allgemein gstoppt und gseit döt ischs falsch oder so. Aso iz einglech nöd usserd da halt das so nöd alles anenand isch sondern so, chli so Pause so
VT	Verbesserungstipps			

RQ1-20: VTFB		Feedback	Tipps, die die SuS der LP spezifisch zur Verbesserung des Feedbacks geben	Ja nur die Lehrerin, Lehrerin könnten noch paar Tipps zu diese Mädchen geben.
RQ1-21: VTSpr		Sprache	Tipps, die die SuS der LP spezifisch zur Verbesserung der Sprachkompetenzen geben	
RQ1-22: VERFB	Vertrautheit Feedback		Aussagen dazu, ob und wie sich die SuS das gehörte FB gewohnt sind oder inwiefern es von dem abweicht, was ihnen vertraut ist.	es war SCHON anders, definitiv, aber nicht dass ich nicht das nie gehört habe oder NIE verstehen, nie verstehen könnte
Feedback als soziales Artifakt zur interaktionalen Mediation von Wissen, Verständnis und Lernfortschritt: Mediationskompetenz				
RQ1-35: IK	Interaktionskompetenz		Als LP mit den SuS angemessen interagieren können und wollen, Gespräche führen vs. Monologe abhalten, dialogischen vs. monologisches Vorgehen	Aso dass man halt beim, wenn man eine Lehrperson ist und den Kindern was beibringen will dass man wie… d/… die Stimmung so zeigt. Also nicht einfach alles so durchredet, dass man auch mit denen sprecht… zusammen spielen kann sozusagen.
ENG	Engagement			
RQ1-25: ENGAllg		Engagement: Allgemein	Aussagen dazu, ob und wie die LP Engagement für die SuS, das Unterrichten und das Fach zeigt, Einschätzungen dazu, ob und wie sich die LP für die SuS einsetzt oder sich für die SuS Zeit nimmt; Aussagen dazu, ob und wie die LP Interesse für die SuS zeigt und den SuS das Gefühl von Wichtigkeit gibt, sie hört und sie wahrnimmt.	und auch für die Mädchen, eeh dass die einfach für sie also ein bisschen Zeit haben sie so zu sagen sie war gut oder NICHT, dann, hat sie ein bisschen beim was hat sie falsch gemacht und NICHT, so jetzt kann sie auf ACHTEN, und so. aber sie war müde, und es t/ und es tönt so dass sie keine… wie soll ich

				das sagen, Interesse dafür hat. […] dass der Person hat keine Intersse dafür einfach, und da fühle mich ein bisschen… SCHLECHT. Ja da/ vielleicht dass ich habe es nicht so gut gemacht oder so.
RQ1-26: ENGMü		Mühe geben	Aussagen dazu, ob und wie sich die LP für die SuS Mühe gibt, sowohl sprachlich als auch didaktisch und pädagogisch. Im Sinne von «put in effort»	es kam halt überzeugend vor, dass sie sich MÜHE gemacht hat. Das… hat mich so überzeugt aso man merkt dass… da Mühe drin steckt und eben auch sich Mühe und sich überlegt hat was sie genau als Rückmeldung sagen will. Ehm ich weiss nicht aber ich denke wei/ aso sie hat sich sich/ au ingwie MÜHE gegeben beim Reden und nicht, einfach, irgendwie geredet
RQ1-27: ENGMot		Motivieren (cf. CEFR-CV, Council of Europe, 2018, p. 106: "The mediation activity 'mediating concepts' involves facilitating and stimulating conditions that are conducive to conceptual exchange and development")	Aussagen dazu, ob und wie die LP die SuS zu ermutigen und motivieren versucht, positiv ist, lobt etc. Prosodische Merkmale sind ausgeschlossen.	Also sie war positiv Si het si AU glaub globt.

RQ1-24: ÜS	Überzeugende Stimme und Ausdruck		Aussagen dazu, ob und wie die LP sprachlich und pädagogisch überzeugend herüberkommt, sei dies aufgrund hoher wahrgenommener Sprachkompetenz oder pädagogischer Kompetenz oder durch den Eindruck, dass die LP den «Ton durchgeben kann».	und, wenn sie, halt, eine überzeugende Stimme dazu hat. Dass… daran merkt ma/ man, dass, s/ s/hat dass sie's kann.
RQ1-30: IR	Inhaltliche Redundanz Feedback		Aussagen zur wahrgenommenen Länge der Sprachproduktion, zu inhaltlichen Wiederholungen, Redundanzen, Schleifen etc.; Aussagen zur «Token Frequency: the total number of words produced»	Es ist einfach, so, auf Dauer halt die g/… gleiche Schleife gewesen. Vielleicht etwas kürzer halten. Sie hat sehr lange umschrieben alles. Mh ich denke sie hätt/ sie GUT verstanden einfach vielleicht irgendwann eben den Faden wie sie vorhin gesagt haben VERLOREN weil sie zu viel gesagt hat.
Forschungsteilnehmende in der Rolle als Beurteilende				
RQ1-31: SBU	Schwierigkeit Beurteilen		Aussagen darüber, wie schwierig es die SuS gefunden haben, die Sprachkompetenzen einer LP zu beurteilen	für mich SEHR schwierig ist, weil ich halt noch ein Schüler bin und noch nicht so hochgebildet bin wie SIE im Englisch. Ehm, da kann ich halt nicht sehr viel sagen dass das einfach für mich ist. So. Es braucht schon… die Zeit um das zu Verstehen mitzubekommen und dann kann ich

				erst die Rückgabe geben. Oder Rückmeldung
RQ1-32: SLPB	Spass LP Beurteilen		Spass am Interview und am Beurteilen von LP	I: wie war das jetzt für dich, Lehrpersonen zu beurteilen? B4: Aso ich fand das toll. I: Ja? B4: Ja. I: Was fandest du toll? B4: Ja dass ich auch mal die Möglichkeit habe zum, aso ja, auch mal Rückmeldungen zu geben.
SON	Sonstiges			
RQ1-33: SONZ		Sonstiges: Zitate		
RQ1-34: SONI		Sonstiges: Interessantes		

Table 47 : Finalised coding frame

SWITCH File

The enclosed USB stick contains the following documents:

General documents:

- Soft copy of the entire dissertation including all appendices in MS Word and PDF format
- Summary

Main-study documents:

- Test specifications and test tasks
- Access to pre- and post-test on Moodle
- Letter of consent for actresses and actors
- Rater training materials:
    - Rating familiarisation task
    - Statement of purpose and rater training
    - Familiarisation rating sheet
    - Rater training meeting (rationale, content, lesson plan)
    - Evaluation rater training
- Rating materials:
    - Rating manual
    - Rating manual benchmarks
    - PRLC-R
    - E-mail with rating update

Sub-study documents (cf. Kuckartz, 2018, p. 222):

- Letter of consent for research participants (field experts)
- Interview guide
- Coding frame with transcription guidelines
- Coded passages that was used for analysis
- All transcripts (raw and annotated)