Language Testing and Evaluation

33

Katrin Wisniewski

Die Validität der Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen

PETER LANG EDITION

Die Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen (GeRS) kommen zunehmend etwa bei der Formulierung von Bildungsstandards oder der Erstellung von Sprachtests zum Einsatz. Dieser Entwicklung steht jedoch ein eklatanter Mangel an Studien zur Möglichkeit der validen Verwendung dieser Skalen gegenüber. Diese Arbeit untersucht die theoretische und empirische Validität der GeRS-Skalen für Wortschatz und Flüssigkeit. Sie zeigt, dass die Skalen auf keiner kohärenten theoretischen Grundlage basieren. Zudem erfassen sie die untersuchte empirische italienische und deutsche gesprochene Lernersprache nur mangelhaft. Die Ergebnisse legen eine umfassende Überarbeitung der GeRS-Skalen nahe, um authentischer Lernersprache und aktuellen Forschungserkenntnissen gerecht zu werden.

Katrin Wisniewski, Studium der Romanistik, Politikwissenschaft, Geschichte und Deutsch als Fremdsprache in Dresden, Bologna und Leipzig; seit 2009 Wissenschaftliche Mitarbeiterin am Institut für Romanistik der TU Dresden; Forschung zur Schnittstelle Sprachtesten/Spracherwerbsforschung, Validitätsforschung, Methodologie.

www.peterlang.com

Die Validität der Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn
and Günther Sigott

Volume 33

Zur Qualitätssicherung und Peer Review der vorliegenden Publikation

Die Qualität der in dieser Reihe erscheinenden Arbeiten wird vor der Publikation durch die Herausgeber der Reihe geprüft.

Notes on the quality assurance and peer review of this publication

Prior to publication, the quality of the work published in this series is reviewed by the editors of the series.

Katrin Wisniewski

Die Validität der Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen

Eine empirische Untersuchung der Flüssigkeits- und Wortschatzskalen des GeRS am Beispiel des Italienischen und des Deutschen

PETER LANG
EDITION

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

Zugl.: Dresden, Techn. Univ., Diss., 2013

Gedruckt auf alterungsbeständigem, säurefreiem Papier.

88
ISSN 1612-815X
ISBN 978-3-631-65015-8 (Print)
E-ISBN 978-3-653-03925-2 (E-Book)
DOI 10.3726/978-3-653-03925-2

Diese Publikation wurde begutachtet.

www.peterlang.com

Danksagung

Die Fertigstellung dieser Arbeit wurde durch den Europäischen Sozialfonds und den Freistaat Sachsen finanziert, wäre aber ohne die fantastische Unterstützung meines Umfelds nicht möglich gewesen. Deshalb möchte ich mich an dieser Stelle ganz besonders bei den Betreuern der Arbeit, Frau Prof. Maria Lieber und Herrn Prof. Erwin Tschirner, bedanken. Ausgesprochen dankbar bin ich auch Herrn Dr. Olaf Bärenfänger, dessen konstruktive Hinweise mir immer wieder eine große Hilfe waren und der stets ein offenes Ohr für schwierige Fragestellungen hatte. Ich danke außerdem ganz herzlich den Kolleginnen der Europäischen Akademie Bozen für ihre Kooperationsbereitschaft und den regen und freundschaftlichen wissenschaftlichen Austausch. Mein Dank gilt hier insbesondere Dr. Andrea Abel, Dr. Chiara Vettori sowie Stefanie Anstein. Ferner möchte ich auch Teresa Knittel, Franziska Plathner und Marzio Piccoli und Sarah Römisch für die Unterstützung danken. Mein besonderer Dank gilt Steffen Horn für sein Verständnis und seinen Beistand, seine konstruktive Kritik und aufbauenden Worte, die diese Arbeit von der Ideenfindung bis zur Abgabe begleiteten.

Inhaltsverzeichnis

Einleitung

Der *Gemeinsame europäische Referenzrahmen für Sprachen* (im Folgenden: GeRS, Europarat 2001a,b, 2004) ist seit seinem Erscheinen im Jahr 2001 zum wohl wichtigsten Bezugspunkt für die Gestaltung von Sprachtests, Bildungsstandards, Curricula, und Lehrbüchern geworden (vgl. etwa Alderson 2007; Fulcher 2004; Hulstijn 2007; Hulstijn/Alderson/Schoonen 2010; Little 2006, 2007; Schneider 2005, 2007). Mit dem Referenzrahmen steht verschiedenen Nutzergruppen ein methodisch komplex aufbereitetes Hilfsmittel zur Beschreibung und Einschätzung kommunikativer Sprachkompetenzen in der Fremdsprache zur Verfügung. Das vom Europarat herausgegebene Dokument hat sich damit als fruchtbares Instrument für die Förderung der europäischen Mehrsprachigkeit herausgestellt. Eine sehr hohe - und stetig steigende - Zahl von Sprachtests (Curricula, Lehrbüchern usw.) wird nunmehr auf den GeRS bezogen, wobei auch wichtige Entscheidungen aufgrund dieser Einschätzungen getroffen werden (so genannte *high stakes*-Tests). Die Gewissheit darüber, dass faire und valide Einschätzungen von Lernersprache mit Hilfe des Referenzrahmens tatsächlich möglich sind, ist deshalb von immenser Bedeutung.

Trotz der weiten Verbreitung des GeRS kann nicht außer Acht gelassen werden, dass das Dokument von Beginn an auch sehr kritisch diskutiert wurde. Zunächst hat man ihm vor allem mangelnde Kohärenz und fehlende Theoriebindung vorgeworfen (vgl. Alderson/Figueras/Kuijper/Nold/Takala/Tardieu 2004; Bausch/Christ/Königs/Krumm 2003; Quetz 2007). Insbesondere in Bezug auf die im Referenzrahmen vorfindlichen Skalen liegt auch ein Mangel an empirischer Validierung vor, der zunehmend angeprangert wird (vgl. Alderson 2007; Fulcher 2004; Hulstijn 2007; Hulstijn/Alderson/Schoonen 2010; Little 2007). Gerade die Skalen und damit das Stufensystem werden aber von vielen Nutzern als essentielles Kennzeichen des GeRS wahrgenommen:

> "Without the scales, the CEFR would have been largely ignored in European language education" (Alderson 2007: 661).

Diese lückenhafte empirische Validierung resultiert aus der Herangehensweise der methodisch durchaus ausdifferenzierten Kalibrierung der GeRS-Skalen im *Schweizer Projekt* des Schweizer Nationalfonds zur Förderung der wissenschaftlichen

Forschung (North 2000; Schneider/North 2000). In einem mehrstufigen Prozess, der hier nur auszugsweise beschrieben werden kann, kategorisierten und rangordneten Bewerter[1] zunächst eine große Zahl an bereits existierenden, aus verschiedenen Testkontexten stammenden englischsprachige Deskriptoren (kurze Beschreibungen von Lernersprache). Die tauglichsten wurden schließlich mit Hilfe des statistischen Verfahrens der Multifacetten-Rasch-Analyse auf einer gemeinsamen Skala der L2[2]-Kompetenz angesiedelt. Dabei fungierten die Entscheidungen der Bewerter faktisch als Daten, während das Rasch-Modell als externer Schiedsrichter wirkte, der über die Tauglichkeit der Deskriptoren entschied (vgl. Fulcher/Davidson/Kemp 2011: 7). Über die entstandenen Skalen weiß man demzufolge, dass sie für die am Projekt beteiligten Bewerter gut handhabbar waren, dass die dort enthaltenen Aspekte der L2-Kompetenz von diesen Bewertern als wichtig erachtet wurden, und welche Deskriptoren sie als Ausdruck einer höheren oder aber weniger ausgeprägten kommunikativen L2-Kompetenz erachten: D.h. die Skalen drücken eine *Bewertungskonvention* aus.

Dies ist sicherlich sehr nützlich. Der Aspekt der Handhabbarkeit von Skalen durch Praktiker darf nicht unterschätzt werden, wohingegen als bedenklich zu betrachten ist, dass momentan allein diese Perspektive das empirische Fundament der GeRS-Skalen bildet. Empirische Validität können GeRS-Skalen nämlich nur dann beanspruchen, wenn die in ihnen enthaltenen Deskriptoren auf wirkliche Lerneräußerungen, den Beschreibungsgegenstand also, überhaupt anwendbar sind (vgl. Alderson 1991). Es ist bislang aber nicht untersucht worden, ob die Skalen einen Bezug zu empirischer, authentischer Lernersprache haben.

Die Mehrzahl der europäischen Bildungs- und Testinstitutionen verlässt sich damit momentan mit teils großer Selbstverständlichkeit auf ein Skalensystem, das einer vollumfänglichen Überprüfung noch harrt und sich in gewisser Weise zu verselbständigt haben scheint. Brian North[3], Hauptverantwortlichen des *Schweizer Projekts*, war durchaus bewusst, dass die GeRS-Skalen einen Bewerterkonsens

---

1 Zu Gunsten der vereinfachten Lesbarkeit wird in dieser Arbeit für den Plural bei Personen generisch das Maskulinum verwendet.

2 In dieser Arbeit wird der Begriff ‚L2' in allgemeiner Weise sowohl für (ungesteuert erworbene) Zweit- als auch für (gesteuert erworbene) Fremdsprachen verwendet. Krashens (1981) terminologische Unterscheidung zwischen ‚Erwerb' und ‚Lernen' wird nicht übernommen. Zum Modus des L2-Erwerbs der Probanden dieser Studie vgl. Kapitel 1.3.4.1.

3 Ich danke Brian North für die freundliche Zuarbeit sämtlicher Quellskalen des GeRS sowie weiterer relevanter Dokumente für diese Arbeit.

repräsentieren und nichts darüber aussagen, was Lerner in Sprachtests tun (North 2000: 71).

Ziel der vorliegenden Arbeit[4] ist deshalb herauszufinden, wie gut und umfassend drei GeRS-Skalen erfassen können, was Lerner in einer typischen Sprachtestsituation tatsächlich tun und wie kohärent diese Skalen in theoretischer Hinsicht sind. In der Studie wird das Verhältnis ausgewählter Niveaubeschreibungen (A2-B2) dieser Skalen (zum Wortschatzspektrum, seiner Beherrschung sowie der Flüssigkeit) zu spracherwerbs- und sprachtesttheoretischen Aspekten und zu empirischer gesprochener italienischer und deutscher Lernersprache untersucht.

In dieser Arbeit werden Skalenvalidierungen bzw. Skalierungen als mehrperspektivische Prozesse verstanden. Laut dem in Kapitel 1.3.2.2 vorgestellten **Drei-Säulen-Ansatz** müssen Deskriptoren auf einem sorgsam ausdefinierten theoretischen Konstrukt basieren bzw. bei einer nachträglichen Skalenvalidierung auf ein solches zurückführbar sein. Diese erste Säule der Skalenvalidität wird hier als durch eine empiriebasierte Vorgehensweise zu ergänzen verstanden, bei der Nachweise für die Passung von Skala und Lernersprache erbracht werden sollten (Säule 2, vgl. Fulcher 1996; Fulcher et al. 2011; Upshur/Turner 1995, 1999). Zudem müssen Skalen auch handhabbar sein, darf die Praktikabilität also nicht außer Acht gelassen werden (Säule 3). Handhabbarkeit wird aber nicht als durch eine gute Inter-Rater-Reliabilität automatisch hergestellt verstanden, sondern hier werden Nachweise dafür als nötig erachtet, dass (auch übereinstimmende) Urteile tatsächlich auf den verwendeten Skalen beruhen.

Die hier untersuchten **Forschungsfragen** (vgl. Kapitel 1.3.1) beziehen sich auf diese drei Säulen der Skalenvalidierung:

Zunächst ist von Interesse, inwiefern die untersuchten Skalen theoretisch kohärent sind, da dies zur Skalenvalidität entscheidend beiträgt (Säule bzw. Forschungsfrage 1). Deshalb werden die Konstrukte der fremdsprachlichen Flüssigkeit sowie des Lernerwortschatzes beleuchtet. Dies dient vor allem dazu, die GeRS-Skalen auf theoretische Bezüge zu analysieren. Die ausführliche Analyse versteht sich jedoch auch als Grundlage für die Ausformulierung möglicher Konstrukte der Flüssigkeit und der lexikalischen Kompetenz in Sprachtests. Vor dem Hintergrund der bislang lückenhaften Modellierung der lexikalischen Kompetenz wird ein Arbeitsmodell dieses Aspekts der kommunikativen L2-Kompetenz entwickelt.

Weiterhin wird das Verhältnis zwischen GeRS-Dekriptoren und authentischer Lernersprache analysiert (Säule bzw. Forschungsfrage 2). Dazu gehört die Untersuchung

4 Die Arbeit wurde von Mai 2010 bis August 2012 vom Europäischen Sozialfonds und dem Freistaat Sachsen gefördert (Projektnr. 80949331).

der Frage, ob die in den Skalen beschriebenen Phänomene von den Lernern tatsächlich (in nennenswerter Anzahl) produziert werden. *Empirische* Beobachtbarkeit und *Relevanz* sind eine Voraussetzung für Validität und beeinflussen außerdem die Reliabilität von Bewerterurteilen. Anschließend wird der Frage nachgegangen, inwiefern die Niveaubeschreibungen dazu geeignet sind, Lernerproduktionen zu erfassen. Unabhängig von Bewertereindrücken wird versucht zu eruieren, ob die alleinige Verwendung der operationalisierten GeRS-Deskriptoren einer Niveaustufe zu sinnvollen Ergebnissen führt und ob die so gruppierten Lerner sich deutlich von anderen Sprechern abgrenzen lassen (*Kriterium der Konvergenz & Trennbarkeit*). Auch wird untersucht, ob die operationalisierten Deskriptoren einen empirischen Bezug zum jeweiligen Skalenkonstrukt aufweisen – damit wird die Frage behandelt, ob die Skalen tatsächlich Flüssigkeit, die Wortschatzbreite und –beherrschung messen, oder ob konstruktirrelevante Aspekte in den Skalen verborgen sind (*Kriterium des empirischen Konstruktbezugs).* Dieser Analyseaspekt liegt an einer Schnittstelle zwischen Theorie und Empirie. Hier wird wiederum das oben erwähnte Arbeitsmodell der lexikalischen Kompetenz vor allem für die Fehlerannotation fruchtbar gemacht.

Neben der theoretischen und der empirischen, lernersprachenbezogenen Validitätsdimension wird auch die Praktikabilität der genannten Niveaubeschreibungen analysiert (Säule bzw. Forschungsfrage 3). Es stellt sich die Frage, ob Bewerterurteile spiegeln, was die Skalen vorgeben, ob sie nachweisbar auf Deskriptoren zurückgehen, und welche anderen Einflussfaktoren aufgezeigt werden können. Hier rückt also das Verhältnis zwischen GeRS-Deskriptoren, Lernersprache und Beurteilungen ins Zentrum.

Die **Datenerhebung** erfolgte im Jahr 2008 im Rahmen des KOLIPSI-Projekts an der Europäischen Akademie Bozen, wo knapp 1.500 Südtiroler Oberschüler auf ihre schriftlichen L2-Kompetenzen getestet wurden (Abel/Vettori/Wisniewski 2012). Einhundert Schüler nahmen am eigens konstruierten mündlichen Sprachtest teil. 19 Schüler (10 Italienisch-, 9 Deutschlerner) wurden aufgrund einer Reihe an Kontrollvariablen für diese Studie ausgewählt. Die Testkonstruktion folgte internationalen Qualitätsstandards (AERA/APA/NCME 1999; ALTE 2001, 2003a-d, 2006 a-d; Bachman/Palmer 1996/2010[5]; Europarat 2009,

5 Bei Bachman/Palmer 2010 handelt es sich um eine überarbeitet Neuauflage von Bachman/Palmer 1996, die zum Zeitpunkt der Testkonstruktion noch nicht erschienen war. Die Neuauflage wurde bezüglich verschiedenster Aspekte der Arbeit ebenso genutzt wie zur Nachkontrolle der Verfahren, die vor ihrem Erscheinen durchgeführt wurden.

2009 [2003][6]; Fulcher 2003; Fulcher/Davidson 2007; Luoma 2004; ausführlicher Abel/Vettori/Wisniewski 2012). Für jede Produktion liegen 4–5 auf Audioaufnahmen basierende, sehr gut übereinstimmende Bewertungen vor, die mit Hilfe eines eng auf den Referenzrahmen bezogenen Bewertungsrasters angefertigt wurden (vgl. Anhang B).

Die Produktionen wurden nach leicht modifizierten CHAT-Transkriptionskonventionen (MacWhinney 2000; vgl. Anhang D) im Annotationseditor ELAN des Max-Planck-Instituts in Nijmegen transkribiert.[7] Anschließend wurden die Deskriptoren so weit wie möglich messbar gemacht und in so genannte ‚Skalenvariablen' übersetzt; diese Operationalisierung war eine Grundlage für die Annotation der Transkripte. Neben diesen GeRS-basierten Indikatoren wurde eine recht große Anzahl verschiedener forschungsrelevanter Aspekte der Flüssigkeit und des Wortschatzes zur Ergänzung der Analysen kodiert (z.B. Strategien oder Verzögerungsphänomene). Die Annotation erfolgte ebenfalls im *multi-layer standoff*-Editor (vgl. Lüdeling/Walter/Kroymann/Adolphs 2005) Elan.[8] Weite Teile der Kodierung wurden zur Erhebung und Kontrolle der Kodierungsreliabilität unabhängig von zwei Kodiererinnen durchgeführt.[9]

Die **Datenaufbereitung und –analse** erfolgte sowohl mit spezifisch sprachbezogener Software (z.B. dem *TreeTagger* vgl. Schmid 1994, oder Wordsmith) als auch mit dem Statistikprogramm SPSS. Es wurde eine Vielzahl den Forschungsfragen angemessener Analysen durchgeführt, darunter neben deskriptiven Verfahren zum Vergleich von Gruppen und Korrelationen vor allem Cluster- und Diskriminanzanalysen.

Die **Ergebnisse** deuten darauf hin, dass mehrere, teils gravierende Aspekte als bedrohlich für die Anwendbarkeit und Validität der drei analysierten Skalen zumindest hinsichtlich der hier untersuchten Lerner zu betrachten sind. Dazu gehört die Tatsache, dass ein Theoriebezug der Skalen in der Regel kaum oder gar nicht herstellbar ist bzw. auch im Text des Referenzrahmens selbst kaum Informationen zu den zu unterstellenden Skalenkonstrukten zu finden sind. Die

---

6 Die Handbuchausgabe von 2009 stand ebenfalls zum Zeitpunkt der Testkonstruktion noch nicht zur Verfügung, wurde aber für verschiedene Aspekte der Qualitätskontrolle (z.B. bei der Datenauswertung) verwendet.

7 Online zugänglich unter http://www.lat-mpi.eu/tools/elan/, Oktober 2013. Die Transkriptionen sind im .txt-Format und innerhalb der Elan-Dateien (.eaf-Format) auf Anfrage bei der Autorin zugänglich.

8 Alle Annotationen und die zugrunde liegenden Audio-Aufnahmen sind auf Anfrage an die Autorin auf DVD frei verfügbar. Die Tagsets finden sich in Anhang C.

9 Ich danke Teresa Knittel für ihre Mitarbeit.

Analyse der Lernersprache ergab zudem, dass die operationalisierten Deskriptoren häufig nicht geeignet waren, um empirisch beobachtbares Lernerverhalten im Korpus zu erfassen; ebenso wurden teils mangelhafte oder gar widersprüchliche Konstruktbezüge offenbar. Auch die Praktikabilität der drei Skalen stellte sich im Einklang mit Ergebnissen der Forschung zur Variabilität im Bewerterverhalten als lückenhaft dar: regelmäßig rekurrierten Bewerter auf skalenexterne bzw. auch konstruktirrelevante Größen.

Die bisher erreichte und zweifellos enorm positive Wirkung des Referenzrahmens auf das Lehren, Lernen und Prüfen von Sprachen auf internationaler, nationaler und regionaler Ebene muss anerkannt werden. Dennoch liefern die Resultate der Untersuchungen Hinweise auf Aspekte mangelnder Validität der GeRS-Skalen. Das Skalensystem scheint sich verselbständigt, in gewisser Weise reifiziert zu haben, ohne dass die Beschaffenheit der Skalen dazu eine ausreichend sichere Grundlage böte. Darauf deuten die Ergebnisse der vorliegenden Studie hin, die sich allerdings aufgrund des Untersuchungsdesigns, d.h. der tiefgehenden Analysen mündlicher Lernersprache, auf wenige Fälle (N=19) beschränken musste und nur sehr vorsichtige Schlüsse erlaubt. Neben den kritisch diskutierten Aspekten werden Möglichkeiten vorgeschlagen, bereits durch kleinere Veränderungen ein vermutlich besseres Funktionieren der Skalen zu erreichen: der Referenzrahmen versteht sich erfreulicherweise als offenes Dokument (vgl. Europarat 2001b: 10).

Die Ergebnisse müssten an einem wesentlich größeren Korpus für mehr Zielsprachen überprüft werden. Erst umfassendere Studien könnten zeigen, wie die anderen GeRS-Skalen und –Niveaus sich für mehr Sprachen und Sprecher im Mündlichen und Schriftlichen bewähren. Ein in diese Richtung zielendes Unternehmen ist das *merlin*-Projekt, das derzeit von der TU Dresden koordiniert wird.[10] Bestätigen sich die hier entwickelten Bedenken, wäre eine umfassende Neu-Skalierung vonnöten, die dann (mindestens) auf der dreifachen Perspektive der Theoriegebundenheit, der Koppelung an empirische Lernersprache und der Berücksichtigung von Aspekten der Handhabbarkeit beruhen könnte.Der momentan beobachtbare Einsatz GeRS-basierter Skalen wäre nicht zu rechtfertigen, wenn die oben genannten Validitätsaspekte sich auch in größerem Rahmen als derartig problematisch erweisen würden.

---

10 Im merlin-Projekt wird eine größere Anzahl schriftlicher Lernerproduktionen des Italienischen, Deutschen und Tschechischen unter anderem auf Indikatoren hin annotiert, die aus Operationalisierungen von GeRS-Skalen stammen. Das Projekt wird im Programm Lebenslanges Lernen der EU von 2012–2014 gefördert (518989-LLP-2011-DE-KA2-KA2MP).

Der **Aufbau** der Arbeit folgt der oben dargestellten Argumentation. Im ersten Kapitel wird zunächst die Problemlage geschildert, indem zunächst grundsätzlich auf Typen sowie Erstellungs- und Analyseverfahren von Skalen im Sprachtestbereich eingegangen wird (Kapitel 1.1). Anschließend werden die Skalen des Referenzrahmens hinsichtlich verschiedener Gesichtspunkte wie etwa ihrer Generalisierbarkeit und ihrer Funktionen kritisch besprochen, woraus ihr Validitätsdefizit deutlich wird (Kapitel 1.2). Im Anschluss wird das Design der Arbeit vorgestellt, wobei zunächst Forschungsfragen und Validierungsansatz entwickelt werden (Kapitel 1.3.1 und 1.3.2), bevor der Fokus auf den verschiedenen verwandten Methoden der Datenerhebung, -aufbereitung und –analyse liegt (Kapitel 1.3.3 und 1.3.4). Auch die Operationalisierung der GeRS-Deskriptoren wird in diesem Abschnitt erläutert.

In den nachfolgenden Kapiteln werden die Forschungsfragen behandelt. Abschnitt zwei widmet sich jeweils zunächst den Konstrukten der L2-Flüssigkeit und der lexikalischen Kompetenz sowie ihrer Messbarkeit, um darauf aufbauend die Frage nach der theoretischen Kohärenz der drei betrachteten Skalen behandeln zu können (erste Forschungsfrage). Das dritte Kapitel ist der empirischen Validierung gewidmet (zweite und dritte Forschungsfrage). Hier wird zunächst auf Aspekte des Bezugs zwischen Lernersprache und den Niveaubeschreibungen der drei in Frage stehenden GeRS-Skalen eingegangen, jeweils anschließend auf den Umgang der Bewerter mit ihnen.[11]

11 Ein umfangreicher Anhang vervollständigt die Arbeit. Weitere Daten (Transkripte, Annotationen, Audiodateien) werden auf Anfrage von der Autorin auf einer DVD zur Verfügung gestellt.

1. Hintergrund

Im folgenden Kapitel wird die Problemlage, die zur Entstehung dieser Arbeit geführt hat, geschildert. Der GeRS soll dabei keineswegs in all seinen Facetten dargestellt werden, sondern es werden nur die für die Fragestellungen dieser Arbeit wichtigen Aspekte beleuchtet. Dadurch rücken aber die vermutlich schwächsten Charakteristika des Dokuments ins Zentrum. Dies soll die vielen positiven Auswirkungen des GeRS keineswegs in Abrede stellen: Wie bereits angedeutet hat das Niveaustufensystem des GeRS enormen Einfluss auf die Bewertung, die Lehre und das Lernen fremder Sprachen in ganz Europa, aber zunehmend auch weltweit (vgl. Tschirner 2012). Erstmals verfügt Europa mit dem Referenzrahmen, dessen Skalen oft als ‚Herzstück' betrachtet werden, über ein Bezugswerk, das die Standardisierung des Lehrens und Prüfens von Sprachen vereinfacht und transparent macht und sehr viel zur Qualitätsentwicklung in diesen Bereichen beigetragen hat:

> "It is our belief that, whatever its shortcomings, the CEFR has introduced a notion of levels of development that is far better – if only because it can be challenged – than the vague terms (not measures) used to date" (Hulstijn et al. 2010: 16).

Der Referenzrahmen ist in 30 Sprachen übersetzt worden, und zahlreiche GeRS-bezogene Initiativen v.a. des Europarats helfen dabei, den Gebrauch des GeRS in vielen Bereichen zu etablieren.[12] So gibt es, um ein Beispiel zu nennen, die so genannten 'Reference Level Descriptions' (RLDs), d.h. einzelsprachspezifische Illustrierungen der GeRS-Niveaus; für das Deutsche existiert seit 2005 die zweite *Auflage von Profile deutsch* (Glaboniat/Müller/Rusch/Schmitz/Wertenschlag 2005). Diese Projekte sind zunehmend korpusgestützt, so z.B. das *Profilo della lingua italiana* (Spinelli/Parizzi 2010), die norwegischen RLD *Norsk Profil* (Carlsen 2013) oder das sehr umfangreich angelegte *English Profile Project* (Hawkins/Filipovíc 2011).[13] Die RLD bemühen sich um eine Veranschaulichung der GeRS-Niveaus, die sie als gegeben voraussetzen und mit vielfältigem Material anreichern.

---

12 Auf der Seite http://www.coe.int/t/dg4/linguistic/Publications_EN.asp (Oktober 2013) stellt der Europarat sehr übersichtlich eine große Anzahl verschiedener (nicht nur) GeRS-bezogener Ressourcen und Publikationen zur Verfügung.

13 www.englishprofile.org/undhttp://www.lanuovaitalia.it/profilo_lingua_italiana/origini.html, Oktober 2013.

Daneben gibt es auch Studien zum Ausrichten von Tests am GeRS (vgl. Figueras/Noijons 2009; Figueras/North/Takala/Verhelst/VanAvermaet 2005), allerdings bislang keinerlei Arbeiten zur empirischen Validität der GeRS-Skalen.

1.1 Skalen zur Beurteilung fremdsprachlicher Kompetenz

Skalen zur Beurteilung fremdsprachlicher Kompetenzen (*proficiency scales* oder *rating scales*[14]) sind seit den 50er Jahren, als die einflussreiche Skala des US-amerikanischen *Foreign Service Institute* (FSI) entstand (vgl. Fulcher/Davidson 2007), weit verbreitet (North 2000: 14; für eine Übersicht Alderson 1991). Sie eignen sich besonders für die subjektive, da von Beurteilern abhängige, Einschätzung produktiver und interaktiver mündlicher und schriftlicher Leistungen in offeneren Testformaten.

Derartige Skalen – auch als „band scores, band scales, profile bands, proficiency levels, proficiency scales, proficiency ratings" (Alderson 1991: 71) bezeichnet – unterscheiden sich teils erheblich, wie in Kapitel 1.1 deutlich wird. Immer wird jedoch versucht, mit ihrer Hilfe (horizontale) Aspekte sprachlicher Kompetenz hierarchisch ansteigend beschreibbar zu machen. Sie messen die Größe, die Häufigkeit, die Intensität, die Bedeutung oder den Rang mit Bezug auf die Tiefe oder die Breite einer demonstrierten Fähigkeit (Hudson 2005: 207).

Skalen betonen positive Eigenschaften der Lernersprache. Harsch (2005: 137) hebt als weitere Stärken hervor, dass skalenbasierte Bewerttungen prototypisches

14 Die schwierige und unterschiedlich gehandhabte begriffliche Trennung zwischen *proficiency – competence – ability* wird durch den deutschen Begriff der ‚Kompetenz' überdeckt. Da unklar ist, ob die 'ability' als Bestandteil der 'competence' betrachtet werden kann, wird im Englischen häufig auf den Begriff der 'proficiency' ausgewichen, der allerdings sehr weit definiert werden kann. Der Hintergrund der Zuordnungsschwierigkeiten liegt in sprachtheoretischen Fragen des Verhältnisses von Kompetenz und Performanz, wie sie seit Chomsky (1965) diskutiert werden (etwa: ist die Kompetenz eine individuell unveränderbare mentale Eigenschaft (und schließt die 'ability' aus, so in kognitiven Modellen), oder ist sie kein absolutes, sondern ein relatives Konzept (und die Performanz ist Teil der Kompetenz), vgl. etwa Lyons 1996). 'Proficiency' kann als zwischen Performanz und Kompetenz angesiedeltes Konzept begriffen werden (Europarat 1994a: 21). In dieser Arbeit wird in Anlehnung an die von Bachman/Palmer 1996 entwickelte, weit verbreitete Terminologie (*communicative language ability*) und an die Definition im GeRS (Europarat 2001b: 109ff.) von ‚kommunikativer Sprachkompetenz' gesprochen (ebenso: Glaboniat 1998).

Verhalten zu beschreiben in der Lage sein können, detaillierte Informationen in Form von Deskriptoren[15] bieten und die Reliabilität von Beurteilungen erhöhen können. Gleichzeitig sind sprachliche Produktionen mit Skalen vergleichbarer, und sie können zu Systemen (wie im Referenzrahmen) zusammengestellt werden (vgl. North 2000: 12). Bachman/Palmer (2010: 352) unterstreichen, dass man mit *rating scales* Informationen über Lernersprache erhalten kann, die auf anderem Wege nur sehr schwierig zu erlangen wären.

Andererseits bergen die auf solchen Skalen basierenden Einschätzungen *Reliabilitätsprobleme*. Bewerterurteile sind durch inkonsistentes Verhalten ein und derselben Person über verschiedene Bewertungen hinweg (mangelnde Intra-Rater-Reliabilität) ebenso bedroht wie durch Unterschiede *zwischen* mehreren Beurteilern (mangelnde Inter-Rater-Reliabilität), bspw. wenn Bewerter verschieden streng beurteilen (Bachman/Palmer 2010: 352). So betont North (1994: 26), dass das Ziel von 'Rating'-Verfahren darin bestehen müsse, die unvermeidliche Subjektivität zu systematisieren. Selbst wenn dies der Fall ist und mithilfe einer Skala reliable Urteile gefällt werden, sie also praktikabel ist, kann ihr jedoch nicht automatisch *Validität* unterstellt werden. Skalen müssen einen nachweisbaren Konstrukt-/Modellbezug haben und empirisch relevant sein (vgl. Kapitel 1.2.3. und 1.3.2).

Ein weiterer Kritikpunkt gegenüber Kompetenzskalen betrifft den häufig fehlenden Nachweis für ihre *Kompatibilität mit Ergebnissen der Fremdsprachenerwerbsforschung* (Turner/Upshur 2002: 51). Konkret wird vielen Skalen zudem vorgeworfen, dass ihre *Deskriptoren* schlecht und unrealistisch gruppiert sowie vage und relativ formuliert sind (Turner/Upshur 2002: 51; Upshur/Turner 1995: 6). Ein Hauptkritikpunkt auch aus der Perspektive dieser Arbeit liegt darin, dass die Deskriptoren nicht oder nicht ausreichend empirisch relevant sind bzw. die auf einem Niveau erwähnten Aspekte nicht in der vorhergesagten Weise *gemeinsam* auftreten (Fulcher 1996; Upshur/Turner 1995).

1.1.1 Arten von Skalen

Skalen zur Einschätzung fremdsprachlicher Kompetenz unterscheiden sich bezüglich einer teils eng zusammenhängenden Reihe von Faktoren. Der folgende knappe Überblick fokussiert für diese Studie zentrale Aspekte.

15 Deskriptoren sind Aussagen in Skalen, mit denen sprachliche Kompetenzen beschrieben werden.

Die wohl wichtigste Differenzierung betrifft die *(1) Generalisierbarkeit* der Aussagen, die mittels einer Skala angestrebt wird:

> "Sampling has to find a defensible way of covering a range of contexts … to enable claims to be made about what sort of performance is likely to be made in untested contexts" (Skehan 1998: 155).

Hierzu ist die Unterscheidung zwischen zwei Skalentypen hilfreich: Skalen, aufgrund derer einschätzbar werden soll, wie ein Proband eine konkrete Test-Aufgabe im ‚wirklichen Leben' zu lösen imstande wäre, werden als 'real-life'-Skalen (Bachman 1990) bzw. aktueller als Skalen bezeichnet, die auf einem '*can do*'-*Ansatz* basieren (Bachman 2011). Diesem Typus rechnet Bachman auch die GeRS-Skalen zu (ebda.). Solche Skalen passen zu dem als 'New Behaviourism' definierten Validitätsverständnis (Chapelle 1998). Ausschlaggebend ist der Testkontext, der so genau wie möglich der Lebenswelt nachgebildet sein muss. Der Generalisierbarkeitsanspruch ist bei solchen Skalen übersichtlich. Es wird nicht unbedingt von einer stabilen, sprecherinternen Kompetenz ausgegangen:

> "Validity would be the degree to which it could be shown that there is a correspondence between the real-world facets and the test facets, and score meaning could only be generalized to corresponding real world tasks" (Fulcher/Davidson 2007: 16).

Skalen hingegen, bei denen aufgrund einer (oder mehrerer) Performanzleistung(en) Aussagen über die Kompetenz getroffen werden, können als '*ability-based*' bezeichnet werden (Bachman/Palmer 2010: 341; der ältere Begriff ist der des 'interactive-ability approach'). *Ability*-basierte Skalen müssen auf ein Modell oder Konstrukt der L2-Kompetenz bezogen sein.[16] Solche Skalen entsprechen einem interaktionistischen Validitätsverständnis (Chapelle 1998: 34): Das Resultat eines Tests wird als Ergebnis von Konstrukten, Kontextfaktoren und deren Interaktion begriffen (vgl. das Arbeitsmodell zur lexikalischen Kompetenz, Kapitel 2.2.2.6).[17] Solche Skalen wollen Aussagen über die Ausprägung bestimmter Bereiche der kommunikativen Kompetenz ermöglichen; sie lassen sich auf 'target

16 Die Skalentypen ('real life' vs 'ability-based') unterscheiden sich aber auch hinsichtlich anderer Faktoren, z.B. bezüglich der möglichen Testinterpretationen. Vgl. für eine Übersicht Lyle Bachmans Vortrag auf der ALTE-Konferenz in Krakau im Juli 2011.

17 Die 'trait theory' hingegen nahm an, dass das zu testende Konstrukt eine Eigenschaft des Testkandidaten und die Kompetenz stabil in diesem repräsentiert sei. Das Testergebnis korrespondiert dann mit dem Konstrukt (vgl. Chapelle 1998). Aktueller wird in der Nachfolge Messicks Validität nicht als Zustand, sondern als Argumentationsprozess begriffen, innerhalb dessen Nachweise für die Angemessenheit von Testinterpretationen erbracht werden müssen (Fulcher/Davidson 2007; Kane 2001; Messick 1989).

language use domains' verallgemeinern (Bachman 2011; vgl. Bachman/Palmer 2010: 60–62). Trotz der demgemäß größeren Generalisierbarkeit von '*ability*'-Skalen dürfen diese nur

> „auf die durch die Bewertung tatsächlich elizitierten Prozesse, Fertigkeiten oder Wissensbestände hin verallgemeinert werden" (Harsch 2005: 151, vgl. Alderson 1991: 80).

Auch ability-basierte Skalen können nicht auf unvorgesehene Funktionen, Nutzer und Kontexte übertragen werden. Die höhere Verallgemeinerbarkeit einer Skala bringt eine geringere Beschreibungstiefe mit sich (Hudson 2005: 208).[18]

Die reliefartige Gegenüberstellung der Skalentypen bzw. die Rückführung aller Versuche der Beschreibung fremdsprachlicher Kompetenz auf entweder eine '*can do*'- oder eine '*ability*'-Ausrichtung[19] bei Bachman (2011) läuft Gefahr, Zwischenformen zu vernachlässigen, wie unten (vgl. Kapitel 1.2.2.1) ausgeführt wird. Zur übersichtlicheren Kurzdarstellung der Hauptskalentypen wird dieser Unterscheidung aber grob gefolgt. Bachman/Palmer (2010) gehen nur noch auf '*ability*'-Skalen ein, während '*can do*'-Skalen generell mehr Forschungsaktivität gewidmet zu werden scheint (vgl. Hudson 2005, vgl. Kapitel 1.1.2).

Ein weiterer wichtiger, eng mit der gerade erläuterten Typologie zusammenhängender Aspekt zur Beschreibung von Skalenarten betrifft die Frage nach dem Einbezug von (2) *Kontext* in Skalen.[20] In verhaltensbasierten Skalen (*can do*) muss

---

18 Kritisch wird diskutiert, inwiefern Skalen als *verhaltensbasiert* (*behavioural*) aufzufassen sind. Sowohl *ability*-basierte als auch die damit häufiger assoziierten *can-do*-Skalen betreffen Verhaltensaspekte (vgl. Europarat 1994a: 32). Beide Typen schließen in der Regel kognitive Aspekte aus (kritisch vgl. Weir 2005a). Das liegt daran, dass kognitive Prozesse schwerer beobachtbar sind, Skalen häufig einen vereinfachten, funktionalen Blickwinkel einnehmen, sich auf Erreichtes (*outcome*) beziehen und somit eher auf Verhalten statt auf kognitive Strukturen und Prozesse konzentrieren (Europarat 1994a: 33).

19 Dazu zählen zum Beispiel die ‚Fertigkeiten und Komponenten'-Ansätze von Lado (1961), Ollers Versuch, die L2-Kompetenz als eindimensional zu beschreiben (Oller 1979), die verschiedenen kommunikativen Modelle (z.B. Canale 1983; Canale/Swain 1980; Bachman 1990; Bachman/Palmer 1996 und 2010), aufgabenbasiertes Testen, wie prominent in Skehans Arbeiten vertreten (z.B. Skehan 2001), oder interaktionale Ansätze (Chapelle 1998; He/Young 1998).

20 Auch die Festlegung der horizontalen Skalendimensionen leitet sich von der Gegenüberstellung *can do* vs *ability-based* ab. Bachman/Palmer (2010: 339) postulieren, dass die horizontalen Dimensionen den Komponenten der Sprachkompetenz zu entsprechen hätten. Harsch wiederum schränkt ein: „Skalen (…), die den Grad der Sprachbeherrschung in Bezug auf bestimmte Aufgaben, die *proficiency* im

der Kontext genau spezifiziert werden, während kompetenzbezogene Skalen (*ability*) keine Kontextabhängigkeit aufweisen sollten (vgl. Brindley 1998). Kontextfreie Skalen sind, wie Hudson (2005: 209f.) bemerkt,

> "(…) terse, efficient, and seemingly straightforward in their application (…) operationalizing terms like "small vocabulary" and "vocabulary of moderate size" clearly becomes normative in nature."

Skalen lassen sich weiterhin danach unterteilen, welche (3) *Funktion* sie haben und welche (4) *Zielgruppe* sie erreichen sollen. In diesem Zusammenhang wird immer wieder Charles Aldersons Aufsatz (1991) zitiert, der als Zielgruppen Bewertende, Testautoren und Nutzer herausarbeitet, die Skalen entweder zur Beurteilung (*rating scale*), zur Berichterstattung (*reporting scale*) oder zur Testerstellung (*construction scale*) verwenden können; bei der Analyse und Erstellung von Skalen ist es von großer Wichtigkeit, diese Aspekte mit zu bedenken.

Weiterhin lassen sich (5) *holistische* von *analytischen* Skalen unterscheiden (Shohamy 1988: 173). Bei letzteren misst und bewertet man verschiedene Aspekte der Kompetenz einzeln. Dadurch erhält der Kandidat ein Profil seiner Fähigkeiten.[21] Bei holistischen Skalen hingegen erfolgt eine einzige Gesamtbewertung. Bachman/Palmer (2010: 339) verurteilen Globalskalen, die eine eindimensionale L2-Kompetenz suggerierten und zu Interpretationsproblemen, Unklarheiten bei der Niveauzuschreibung und uneindeutiger Gewichtung von Komponenten durch die Bewerter führen könnten, und empfehlen die ausschließliche Verwendung analytischer Skalen. Holistische Skalen verleiteten außerdem dazu, gute Inter-Rater-Reliabilitätswerte als Validitätsargument umzudeuten (Fulcher/Davidson 2007: 97; vgl. Weigle 2002: 121).

---

interactive-ability approach, beschreiben sollen, müssen mangels einer empirisch validierten Theorie der *proficiency* alle Aspekte der *proficiency* unter Zuhilfenahme des gesunden Menschenverstandes und pragmatisch gehaltener Beschreibungen beachten" (Harsch 2005: 144). Sicher hat Harsch mit ihren Vorbehalten recht; andererseits ist es möglich, Konstrukte aus nur teilweise validierten Theorien zu entwickeln und als Grundlage zur Skalenentwicklung zu verwenden – es ist fraglich, wann (und ob) die Theoriebildung auch nur als vorläufig abgeschlossen betrachtet werden kann und sicherlich nicht wünschenswert, deshalb grundsätzlich auf kompetenzbasierte Skalenentwicklung zu verzichten (vgl. Jones/Saville 2009a; Kane 2001).

21 Zu Vorteilen analytischer Skalen vgl. Bachman/Palmer 1996: 210 und dies. 2010: 338f. Von Nachteil (gerade bei mündlichen Tests) ist allerdings die große Anzahl von Entscheidungen, die Beurteiler treffen müssen, in deren Folge es etwa zum so genannten Halo-Effekt kommen kann (vgl. Fulcher 2003:89; Weir 2005b: 193).

Ferner unterscheiden sich Skalen hinsichtlich (6) des *Messniveaus* (Nominalskalen, Ordinalskalen, Intervallskalen), wobei beim Sprachtesten wohl maximal Ordinalskalenniveau erreicht werden kann (vgl. Europarat 1994a: 17).

Man kann außerdem zwischen (7) *kriteriums- und normorientierten* Skalen unterscheiden: wird eine sprachliche Leistung auf eine konkrete Gruppe bezogen, spricht man von normorientierter Bewertung, liegt der Bewertung hingegen ein externer Standard zugrunde, handelt es sich um ein kriteriumsorientiertes Vorgehen (vgl. z.B. Bachman/Palmer 2010: 342; Europarat 1994a: 27). Brian North (Europarat 1994a: 27–30) arbeitet Bedingungen für kriteriumsorientierte Skalen heraus. Sie dürfen nicht aus relativen Niveaustufen (die nur in Bezug aufeinander Sinn ergeben) bestehen (Bachman 1990: 343–344) und die Deskriptorformulierungen dürfen nicht relativ sein. Außerdem verweist North auf die Gefahr einer Zirkelargumentation, die in dieser Arbeit den (von North später mit entwickelten) GeRS-Skalen zum Vorwurf gemacht wird:

> "Presumably it [the criterion description, K.W.] loses validity if samples or research indicate that what actually happens is different from what is described in the descriptor" (Europarat 1994a: 29).

Als Standards bzw. Kriterien können verschiedene Referenzen dienen, z.B. Niveaustufen (vgl. Bachman/Palmer, ebda.). Eine (8) *Anschlussmöglichkeit* an die *L2-Erwerbsforschung* besteht:

> "But there is a fourth sense of criterion-referenced measurement which is even more difficult to achieve and which has eluded language testers so far. This is that the *proficiency levels which are the basis of criterion referencing are linked in some cumulative way to a course of development*. This would allow each intermediate step simultaneously to have a proficiency, real world dimension, and also a relationship with other stages of development." (Skehan 1989: 6)

Das Verhältnis von Bewertungsskalen zur L2-Erwerbsforschung ist jedoch umstritten. Es ist nicht klar, ob man überhaupt von Erwerbssequenzen sprechen kann oder eher von individuellen, hochvariablen Mustern beim Erlernen von Sprachen (vgl. z.B. Bardovi-Harlig 2006: '*main routes*' vs '*individual paths*'). Während einige Linguisten das Verhältnis von Bewertungsskalen und L2-Erwerb sehr kritisch sehen (vgl. dazu Europarat 1994a: 26), fordert z.B. De Jong (1988: 74):

> "What we need to know if we want to develop good scales is not linguistic knowledge of how language is structured, what all the features of language are; we need to know how somebody acquires language, that is, what the developmental stages in language acquisition are."

Die Erfassung sprachlicher Kompetenzen in Bewertungsskalen ist etwas anderes als die Beschreibung des Erwerbsverlaufs einer L2; jedoch ist beides miteinander verbunden. *Widersprechen* Skalen Erkenntnissen der Erwerbsforschung, ist ihre *Aussagekraft bedroht*; nehmen sie *nicht Bezug* auf die Forschung, laufen sie Gefahr, *irrelevante Aspekte* der L2-Kompetenz abzubilden. Problematisch ist, dass Bewertungsskalen eingesetzt werden (müssen), bevor eine auch nur annähernd vollständige Theorie des L2-Erwerbs vorliegen kann. Dies ist jedoch keine Rechtfertigung für eine Entkoppelung beider Bereiche. Forscher bspw. im SLATE-Netzwerk (*Second Language Acquisition and Testing in Europe*) befassen sich mit besonderem Fokus auf den GeRS auch mit der Frage, wie Erwerbssequenzen mit der kommunikativen L2-Kompetenz zusammenhängen (vgl. z.B. Bartning/Martin/Vedder 2010).[22]

1.1.2 Skalierungsverfahren

Die Konstruktion von Skalen ist ein komplexer Prozess, der eng mit der oben beschriebenen Art der Skala zusammenhängt. Häufig werden keine Informationen zur Skalierung zur Verfügung gestellt (Brindley 1998: 117; Knoch 2011: 81; North 2000: 3). Hudson (2005) diagnostiziert einen Forschungsschwerpunkt bei der Entwicklung von Skalen zum Testen sprachlicher Performanz, wobei die Erledigung einer möglichst authentischen, gründlich durch jeweilig relevanten Kontext definierten Aufgabe im Vordergrund steht (‘*can do*‘-Ansatz).

Bei jedem Skalierungsverfahren müssen eine *horizontale* und eine *vertikale Dimension* bedacht werden. So sollte theoriegestützt (vgl. kritisch Knoch 2011) unter Berufung auf Modelle der kommunikativen Sprachkompetenz (z.B. Bachman 1990; Bachman/Palmer 1996/2010; Canale/Swain 1983) festgelegt werden, welche Aspekte der sprachlichen Kompetenz in der Skala erfasst werden. Außerdem ist eine Entscheidung über die vertikale Dimension, d.h. die unterschiedliche Ausprägung dieser Aspekte, zu treffen. Beides hängt eng mit dem Zweck bzw. der Zielgruppe der Skala zusammen (vgl. Alderson 1991; Harsch 2005).

Hauptsächlich kann man *intuitive* Methoden, bei denen Expertenurteile eine entscheidende Rolle spielen, von *empirisch* gestützten Verfahren unterscheiden.[23]

---

22 http://www.slate.eu.org/index.htm, Oktober 2013.

23 Ganz ähnlich funktioniert Widdowsons (1990) Unterscheidung zwischen ‘type-’ und ‚token-basierten‘ Skalierungen. Fulcher (2003: 88ff.) trennt vier Typen und geht insbesondere auf die Skalen der FSI-Familie ein. North (1994: 38ff.) unterscheidet intuitive, empirische (d.h. quantitative, teilweise empirische) und qualitative (empirische)

Die meisten Skalierungen beruhen auf nur in Ausnahmefällen transparenten und oft *ex post* validierten intuitiven Experteneinschätzungen (Europarat 1994a: 38).

Werden diese mit einem Messmodell auf einer Skala angeordnet (wie beim GeRS, vgl. Kapitel 1.2.1), sprechen Fulcher et al. (2011: 5) von einem '*measurement-driven approach*'. Solche Skalierungsverfahren sind nur insofern empirisch, als Expertenurteile einfließen. Die Autoren kritisieren, dass die Verfahren keine angemessenen Beschreibungen von Lernersprache lieferten und nicht den kommunikativen Kontext sowie die interaktionale Komplexität des Sprachgebrauchs berücksichtigten, da sie sehr abstrakt seien: am Ende sei nicht klar, was eine Bewertung wirklich bedeute (Fulcher et al. 2011: 8f.). Außerdem sei der Ansatz, Bewerterurteile auf einer Rasch-Skala anzuordnen, wie es bei der GeRS-Skala geschah, nicht theoriebasiert (Fulcher 2004: 258). Der Einsatz *probabilistischer* statistischer Verfahren bietet aber auch viele Vorteile (vgl. North 2000, 2007a).

Die empirische Skalenentwicklung ('*performance data-driven*', dies.: 5) hingegen basiert auf Analysen von Lernerproduktionen. Empirische, auf Lernersprache beruhende Skalierungsverfahren sind als Reaktion auf das Validitätsdefizit statistisch skalierter bzw. auf Expertenurteilen beruhender Skalen seit den 90er Jahren vermehrt eingefordert worden (vgl. Fulcher 1996, 2003; Turner 2000; Turner/Upshur 2002; Upshur/Turner 1995, 1999). Hier wurde kritisiert, dass viele Skalen eine nur scheinbar kontinuierliche Progression suggerierten, die von der L2-Forschung nicht bestätigt werden könne (vgl. Bachman/Cohen 1998). In verschiedenen innovativen Herangehensweisen werden alternative Skalierungsmethoden verfolgt, so bei den 'binary-choice boundary definition scales' (EBB) von Upshur/Turner (1995)[24] oder den 'performance decision trees' (PDT) von

---

Methoden. Die Begrifflichkeiten sind anders als bei Fulcher et al. (2011), zielen aber im Grunde auf dieselben Phänomene. Unter quantitativen Methoden versteht North etwa auf klassischer Statistik beruhende Verfahren, Multitrait-Multimethod-Prozeduren (MTMM) oder das mittlerweile häufig eingesetzte Rasch-Verfahren (1994: 40ff.). Auch qualitative Methoden sind in Norths vielleicht etwas sperrigerer Taxonomie teils empirisch; zu ihnen gehört etwa die Schlüsselwortmethode, um wichtige Konzepte für Bewerter oder in den Lernerdaten herauszufinden. Intuitive Methoden sind etwa der Bezug auf Expertenurteile, Expertenkommittees oder die Entwicklung von Skalen aus einem internen Konsens heraus, also nicht auf wissenschaftlichen Säulen beruhende Verfahren.

24 Bei EBB können eine hierarchische Deskriptorenanordnung undAnnahmen über einen linearen Fremdsprachenerwerbsverlauf vermieden werden. Die Bewertung erfolgt in einer Serie von Ja/Nein-Entscheidungen. Diese Skalen sind bewerterorientiert und haben einen 'real-life'-Fokus. Bei EBB-Skalen wird nicht das Prototypische einer Niveaustufe in den Vordergrund gerückt, sondern die Kategoriengrenzen (Upshur/

Fulcher et al. (2011).[25] In diesem Kontext wurde auch die Bedeutung des Einbezugs eines angemessenen Samples an Lernerproduktionen deutlich (vgl. Turner/Upshur 2002).[26]

Trotz der größeren empirischen Validität solcher Skalen liegen Nachteile in der extrem aufwändigen Erstellung sowie in der eigentlich unmöglichen Übertragbarkeit der Skalen auf andere Test- und Bewertungskontexte (z.B. Turner/Upshur 2002: 53). Widdowson (1990: 75) ist der Auffassung, dass nicht in jedem Fall ein empirisches Skalenentwicklungsverfahren einem intuitiven vorzuziehen sei. Problematisch ist auch, dass die Versuche empirischer Skalierungen bislang offenbar ausschließlich auf 'real life'-Skalen bzw. Tests konkreter performanzbezogener Fähigkeiten abzielen (wie in Kapitel 4 des GeRS), nicht jedoch auf Versuche, Aspekte der sprachlichen Kompetenz zu erfassen (wie in Kapitel 5 des GeRS).

Auch der Einbezug empirischer Sprachdaten schützt nicht automatisch vor einer fehlgeleiteten Skalierung. Wenn diese nicht durch theoretische Analysen gestützt wird, ist fraglich, ob Skalen erfassen, was besonders relevant für die kommunikative Kompetenz bzw. für den Fremdsprachenerwerb ist. Ein theoretisch belastbares Konstrukt ist für eine valide Skala ebenso essentiell wie ein robuster Bezug auf Lernersprache und die Handhabbarkeit durch Bewerter (vgl. Kapitel 1.3.2).

Knoch (2011) schlägt eine *theoriegeleitete Skalenentwicklung* (für das Schreiben) vor. Auch wenn vorhandene Theorien nach Ansicht der Autorin für die Skalenerstellung eigentlich nicht ausreichen, müsse möglichst theoriebasiert

---

Turner 1995: 10). Die Skalen sind aufgabenbasiert, so dass die Ergebnisse nicht verallgemeinert werden können. EBB-Skalen beruhen auf Bewertereindrücken: Bewerter sortieren (ohne Skala) Lernerproduktionen in eine gegebene Anzahl an Niveaustufen und legen dann gemeinsam Kennkriterien dieser Niveaus fest.

25 Fulcher et al. (2011) reichern in ihren PDT das EBB- Bewertungskonzept durch eine gründliche Beschreibung der Test-Domäne bzw. des genauen kommunikativen Kontexts an und entwickeln einen Beispiel-Entscheidungsbaum für Verkaufsgespräche in Reisebüros. PDTs kommen ohne Annahmen über die hierarchische Ausprägung von Deskriptoren aus; ihre Validität hängt allerdings direkt davon ab, wie angemessen die Beschreibungen der Sprachdomäne sind, auf denen sie beruhen (dies.: 23).

26 Turner/Upshur (2002) untersuchen, inwiefern verschiedene Entwicklerteams solcher EBB (zum Schreiben) sowie verschiedene empirische Lernerdaten, die diesen bei ihrer Arbeit zur Verfügung standen, zu denselben Skalierungsergebnissen führen. Es stellte sich heraus, dass unterschiedliche Skalenentwicklungsteams zu sehr ähnlichen Skalen kamen, jedoch unterschiedliche zugrunde liegende Samples zu ganz anderen Skalen führten (dies.: 65). Die Auswahl der Lernerproduktionen, auf deren Grundlage eine Skalierung erfolgt, ist also von entscheidender Bedeutung für die Charakteristik empirischer Skalen.

vorgegangen werden (vgl. Jones/Saville 2009a). Knoch empfiehlt eine Taxonomie, die auch Performanzaspekten Rechnung trägt. Ihr Hinweis darauf, dass in Modellen der kommunikativen Sprachkompetenz der Performanz keine Aufmerksamkeit zuteil wird, obwohl gerade Aspekte wie die Flüssigkeit für die Bewertungspraxis von großer Bedeutung sind, ist sehr bedeutsam. Die von ihr vorgeschlagene Taxonomie hat allerdings keinen Modellcharakter, da die einzelnen Bestandteile nicht aufeinander bezogen werden, so dass fraglich ist, inwiefern sie bei einer Skalierung hilfreich sein können.

1.1.3 Die Analyse von Skalen

Betrachtet man Validität als Argumentation für die Angemessenheit von Testinterpretationen (vgl. Bachman/Palmer 2010; Chapelle 1999; Kane 2001, 2013; Messick 1989), müssen Skalen als wichtiger Einflussfaktor auf sie aufgefasst werden. Aufgrund von auf Skalen beruhenden Bewerterurteilen werden 'scores' für Testleistungen berechnet, deren Interpretationen valide sein sollen. Sind die verwendeten Messinstrumente aber empirisch nicht relevant, theoretisch nicht unterfüttert und praktisch nicht handhabbar (vgl. Drei-Säulen-Konzept, Kapitel 1.3.2.2), ist es so gut wie ausgeschlossen, zu validen Interpretationen von Testergebnissen zu gelangen. Deshalb müssen auch Skalen selbst validiert werden, im Idealfall natürlich bereits bei ihrer Erstellung. Dies scheint jedoch der Ausnahmefall zu sein, geschieht zudem häufig *ex post* und selten in transparenter Form (Knoch 2007). Die Analyse von Skalen, die selten explizit zum Zweck der Validierung erfolgt, ist in der Sprachtestforschung ein Thema von geringer Bedeutung. Obwohl Skalen die Validität von Testinterpretationen erheblich beeinflussen und teils als ‚Quasi-Konstrukte' fungieren (vgl. Knoch 2007; McNamara/Hill/May 2002), befassen sich Validierungsstudien selten direkt bzw. ausschließlich mit Skalen. Dies ist auch darauf zurückzuführen, dass Skalen i.d.R. in einen konkreten Testkontext eingebettet sind, der dann umfassender auf Validitätsaspekte beleuchtet wird.

Diese Arbeit geht davon auss, dass für die Möglichkeit eines validen Einsatzes von Skalen mindestens Nachweise dafür zu erbringen sind, dass die betreffende Skala auf Theorien oder Modelle fremdsprachlicher Kompetenz bezogen und für Beurteilende handhabbar ist sowie Charakteristika empirischer Lernersprache spiegelt (vgl. Kapitel 1.3.2). Diesen Aspekten wird in der Literatur in ungleichem Maße Aufmerksamkeit zuteil: Die überwältigende Mehrheit der Studien misst den Beurteilern und damit Fragen der Praktikabilität mehr oder weniger direkt eine zentrale Rolle zu. Im Folgenden wird zunächst diese Gruppe von Arbeiten vorgestellt. Im Anschluss werden einzelne Skalenanalysen präsentiert, die von diesem Muster abweichen.

Bei der Gruppe der *bewertungsfokussierten Studien* geht es einerseits häufig etwa darum herauszufinden, welche Kriterien Bewerter bei ihrer Arbeit als besonders relevant empfinden, um diese dann zu skalieren (vgl. die Übersicht bei McNamara/Hill/May 2002: 229ff.). Brown/Iwashita/McNamara (2005) z.B. arbeiten von Bewertern verwandte Beurteilungsaspekte heraus. Sie zeigen, dass viele dieser Aspekte zwischen (bewerteten) Sprachkompetenzniveaus differenzieren können (dies.: 82). Iwashita/Brown/McNamara/O'hagan (2008) untersuchen anhand einer fünfstufigen Globalskala im Vorhinein bewertete Testleistungen von 200 LernerInnen des Englischen zur Beantwortung der Frage, ob und hinsichtlich welcher Kriterien sich die Sprache auf den verschiedenen Niveaus unterscheidet. Hier erwiesen sich insbesondere die Messgrößen zum Wortschatz und zur Flüssigkeit als niveaudifferenzierend (dies.: 42). Diese hatten den größten Einfluss auf das Gesamtergebnis der Kandidaten (dies.: 41).[27]

In anderen Studien fließen Bewerterurteile weniger explizit in Skalenanalysen mit ein, etwa wenn Skalenniveaus als gegeben und Ausgangspunkt für weitere Analysen betrachtet werden. So ist es etwa Ziel des *English Profile Project* (vgl. Hawkins/Buttery 2009, 2010; Hawkins/Filipović 2011), in der Tradition der vom Europarat initiierten *Reference Level Descriptions* (s.o.) eine auf umfangreichen Sprachtest-Korpora basierende Illustration der GeRS-Skalenniveaus zu liefern und typische Phänomene ('*criterial features*') für die einzelnen Niveaustufen herauszufiltern:[28]

> "The result is, however, that the CEFR levels are underspecified *with respect to key properties that examiners look for* when they assign candidates to a particular proficiency level and score in a particular L2 (…). The basic intuition behind the criterial feature concept is that there are certain linguistic properties that are characteristic and indicative of L2 proficiency at each level, *on the basis of which examiners make their practical assessments*." (Hawkins/Buttery 2010: 2; Hervorhebungen K.W.).

---

27 Allerdings wurden kleine Effektgrößen und große Standardabweichungen gemessen, "indicating broad variation among learners at any one level, and overlap between levels" (Iwashita et al. 2008: 41). Die Ergebnisse dieser Studie scheinen Annahmen zu widersprechen, laut denen Bewerter sich bei ihrer Arbeit vor allem auf die grammatische Korrektheit konzentrieren, unterstreichen aber auch die Vielfalt an Einflussfaktoren auf die Urteile.

28 Auch im bereits erwähnten SLATE-Netzwerk werden solche typischen Phänomene für einzelne GeRS-Niveaustufen gesucht (vgl. Forsberg/Bartning 2010), allerdings wird dort auch erforscht, inwiefern (auf Beurteilungen beruhende) Skalenniveaus u.a. des GeRS an Erkenntnisse der Spracherwerbsforschung ankoppelbar sind (Bartning et al. 2010).

Im *English Profile Project* wird also der Ansatz des Referenzrahmens, in dem die Niveaustufen ausschließlich durch Bewertereindrücke gerechtfertigt sind, im Grundsatz übernommen.

Studien, die in der beschriebenen oder in ähnlicher Weise auf Bewertereindrücke rekurrieren, dürfen nicht als Validitätsnachweise für Skalen missverstanden werden. Es wird nämlich *erstens* meist die Existenz gegebener Niveaustufen präsupponiert und nicht weiter hinterfragt, insbesondere wenn eine hohe Inter-Rater-Reliabilität vorliegt (vgl. Connor-Linton 1995; Cumming 1990; Lumley 2002; Weigle 1994). In dieser Argumentation können aber lediglich die Reliabilität und Grundlage von Bewerterurteilen eingeschätzt werden, nicht die Validität der Skala selbst; *Reliabilität und Validität werden verwechselt* (vgl. ausführlicher dazu Kapitel 1.3.2.2). Reliabilität kann als notwendige, aber nicht hinreichende Bedingung für die Validität (vgl. aber Moss 1994) bzw. als Aspekt der Validität (vgl. z.B. Weir 2005a) gelten.[29]

Zweitens darf die bekannte *Variabilität* des Verhaltens auch geschulter Bewerter nicht außer Acht gelassen werden. In den auf Praktikabilität fokussierenden Studien wird der Zusammenhang zwischen Lernersprache und Skalenniveaus ja nur *vermittelt* hergestellt: Das idealiter theoriebasierte *Konstrukt* von L2-Kompetenz findet sich – so ist zu hoffen – implizit in den *Skalen*. Diese wiederum gehen durch die Hände von *Bewertern* – die eventuell aber von den Skalen und damit auch dem Konstrukt abweichende Charakteristika der produzierten L2 bewerten (vgl. Connor-Linton 1995: 763, vgl. ausführlicher Kapitel 1.3.2.2).

Drittens besteht die Gefahr eines logischen Zirkelschlusses (vgl. Hulstijn 2010: 235), wenn bei einer Skalenvalidierung bewertete Niveaustufen als Ausgangs- und Endpunkt verwendet werden: man versucht dann, die Existenz von etwas zu beweisen, das man bereits als gegeben hingenommen hat.

Alternative, nicht (ausschließlich) auf Bewertereindrücke rekurrierende **Ansätze** legen Ute Knoch und Claudia Harsch vor. Ute Knoch bezieht die Validität von *rating scales* auf die Kategorien zur Beschreibung der Nützlichkeit von Tests nach Bachman/Palmer (1996) (2009a: 64f., vgl. Tab. 1). Ihre Skalenvalidierung orientiert sich zudem am *assessment use argument* von Bachman (2005).

---

29 Das Missverständnis wird auch bei McNamara et al. (2002: 8) deutlich: "Although the number of studies investigating test-taker discourse has been growing, to date few studies have examined the relationship between test scores and the substance of the performance on which it is based in order to validate the rating scales used in an assessment."

Tab. 1: Knochs Aspekte der Validität von Bewertungs-Skalen (2009: 65)

Aspekt der Nützlichkeit	Anforderungen an valide Skala
Konstruktvalidität	Die Skala führt zu dem beabsichtigen, zweck- und kontextangemessenen Outcome und Bewerter nehmen die Skala als konstruktangemessen wahr. Die Skala unterscheidet gut zwischen Probanden und die Bewerter berichten, dass die Skala gut funktioniert Die Deskriptoren reflektieren aktuelle Erkenntnisse der Angewandten Linguistik ebenso wie der Forschung [sic]
Reliabilität	Bewertungen sind reliabel und austauschbar.
Authentizität (Inhaltsvalidität)	Skalen reflektieren so gut wie möglich, wie Schreiben von Lesern in der TLU-Domäne wahrgenommen wird.
Impact	Das Feedback für die Probanden ist relevant, vollständig und bedeutsam Testscores und Feedback werden auch von anderen Stakeholdern als relevant, vollständig und bedeutsam wahrgenommen Impact auf Rater ist positiv
Praktikabilität	Der Skalengebrauch ist praktikabel [sic] Skalenentwicklung ist praktikabel.

Tab. 1 führt vor Augen, dass der in dieser Arbeit als ‚Praktikabilität' bezeichneten Eignung von Skalen für Bewerter bei Knoch eine zentrale Rolle beigemessen wird, und zwar auch bei Aspekten der Konstruktvalidität. Knochs Ansatz zielt auch auf die Wirkung von Skalen ('Impact'), ein wichtiger Punkt, der in der vorliegenden Studie nicht mit betrachtet werden kann (vgl. Jones/Saville 2009b). Kritisch kann auf Begriffsebene zwar angemerkt werden, dass die Definition der Praktikabilität bei Knoch selbstreferentiell ist. Insgesamt findet man jedoch in der Literatur kaum Beiträge, die wie Knochs Arbeiten die Bedeutung der Validität und der Validierung von Bewertungsskalen so explizit machen und auch konkrete Realisierungen vornehmen.[30]

Eine umfassende Skalenanalyse findet sich auch bei Claudia Harsch (2005). In ihrer Dissertation befasst sich Harsch mit der Beurteilung des semikreativen

30 KNOCH (2009a) entwickelt und validiert (2007) eine empirisch und theoretisch fundierte Skala. In innovativer Herangehensweise untersucht KNOCH (2009b) die Validität einer Skala, die in der Luftfahrt eingesetzt wird und betont die Bedeutung des Einbezugs von Stakeholdern.

Schreibens im Projekt *Deutsch-englische-Schülerleistungen international* (DESI)[31] und der Bedeutung des GeRS für diesen Kontext; sie diskutiert zahlreiche Aspekte des GeRS kritisch, analysiert jedoch nicht empirisch die Validität von dessen Skalen. Sie betont, dass es kein

> „objektives Außenkriterium [gibt], an dem Skalen validiert werden könnten. Denn das jeweilige Modell, nach welchem Sprache beschrieben wird, die Merkmale, welche als relevant für bestimmte Aspekte und Niveaus betrachtet werden, oder auch die Sprache der Deskriptoren – all diese Aspekte haben keinen Absolutheitsanspruch, sondern ihre Gültigkeit ist relativ: relativ in Bezug auf das jeweils gültige Paradigma, relativ bezüglich der jeweils vorherrschenden Vorstellung von dem, was beispielsweise Sprachvermögen ausmacht, relativ bezogen etwa auf den Sprachgebrauch der Skalenkonstrukteure“ (Harsch 2005: 149).

Als Qualitätsmerkmale von Skalen postuliert Harsch die Angemessenheit für den jeweiligen Beschreibungsgegenstand sowie die Stimmigkeit der horizontalen (Komponenten der Sprachkompetenz) und vertikalen (Ausprägungen dieser Komponenten) Dimensionen. Außerdem sei entsprechend neueren Erkenntnissen in der (Sprachtest-)Validitätsforschung essentiell, dass Skalen so eingesetzt werden, wie die intendierte Funktion es vorsieht (dies.: 149ff.). Harsch stellt damit Aspekte der Konstruktvalidität in den Vordergrund, während die empirische Relevanz und die Praktikabilität eine untergeordnete Rolle spielen (vgl. Drei-Säulen-Modell, Kapitel 1.3.2.2).

Eine zunehmende Anzahl an Studien befasst sich mit diskursanalytischen *empirischen Analysen von Testsprache,* ohne dabei einen Bezug zu den jeweils verwendeten Skalen herzustellen (für einen Überblick vgl. McNamara et al. 2002). Häufig geht es dabei um den mehr oder weniger konversationellen Charakter von Interviews (He/Young 1998), Methodeneffekte wie den Einfluss des Interviewerverhaltens auf die Testsprache (Brown 2003) oder Einflussgrößen wie die Planungszeit (Wigglesworth 1997, 2001).

Fast alle zitierten empirischen Arbeiten untersuchen das Englische als L2, so dass zur mangelhaften Erforschung der Validität von Skalen im Allgemeinen und der GeRS-Skalen im Besonderen generell eine eklatant einseitige sprachliche Fokussierung tritt.

---

31 Informationen auf der Website des leitenden Deutschen Instituts für internationale pädagogische Forschung, http://www.dipf.de/de/projekte/deutsch-englisch-schueler leistungen-international, Oktober 2013.

1.2 Die Skalen des Gemeinsamen europäischen Referenzrahmens

In diesem Kapitel wird zunächst die Erstellung der GeRS-Skalen beschrieben (Kapitel 1.2.1), die weit reichenden Einfluss auf die Möglichkeiten des Einsatzes der Skalen hat und ihre Validität deutlich einschränkt (Kapitel 1.2.3; vgl. Fulcher 2003, 2004, 2008; Fulcher/Davidson 2007; Fulcher et al. 2011; Hulstijn 2007; Weir 2005b; Wisniewski 2010a, [1], [2] im Druck). Es soll unterstrichen werden, dass der *Text* des Referenzrahmens in einem Prozess entstanden ist, der u.a. auf einflussreiche Initiativen zur Kompetenzbeschreibung als dem, was ein Lerner tun kann (Kannbeschreibungen), zurückgeht (z.B. den *Threshold Level*, Europarat 1975) und von der Skalierung der Deskriptoren selbst *getrennt* verlief (Vogt 2011).

1.2.1 Skalierungsmethode der GeRS-Skalen: Praktikabilität als Leitkriterium

Die methodisch aufwändige und neuartige Skalierung der Skalen des Referenzrahmens fand zwischen 1993 und 1996 im so genannten *Schweizer Projekt* des Schweizer Nationalfonds zur Förderung der wissenschaftlichen Forschung statt und wird detailliert in North (2000) sowie in Schneider/North (2000) dargestellt; eine kurze Zusammenfassung findet sich auch im GeRS selbst (Europarat 2001b: 210–217). Im Folgenden sollen aus Platzgründen nur die wichtigsten Schritte der Kalibrierung referiert werden.

An dem (mehrstufigen) groß angelegten Projekt für die Zielsprachen Englisch, Französisch und Deutsch nahmen ca. 2.800 Lerner und 300 LehrerInnen aus der Schweiz teil. Ziel des Projekts war es, Deskriptoren zu skalieren, die in Verbindung zu den bereits existierenden Kategorien des GeRS stehen sollten, auf kollektiver Erfahrung aufbauen, für Praktiker verständlich sein und mit einem Messmodell kalibriert werden sollten (North 2000: 3).

Dabei wurden intuitive und empirische Phasen durchlaufen. Zunächst wurde eine Sammlung von (ca. 2000) Deskriptoren in Form von Kann-Beschreibungen aus bereits bestehenden Skalen angelegt und analysiert (vgl. Europarat 2001b: 217). Die Deskriptoren entstammten ganz verschiedenen mündlichen Sprachtests (des Englischen, vgl. Schneider/North (2000: 231f.); eine Diskussion einiger dieser Skalen erschien bereits in Europarat (1994). Darunter sind kurze oder ausführliche holistische Quellskalen, die die Sprachkompetenz allgemein beschreiben

sollen (wie etwa die zur *Foreign Service Institute* (FSI)- und *International Language Testing Service* (ILTS)-Familie gehörigen, vgl. North 2000: 22), aber auch Skalen, die sich auf verschiedene Gebrauchskontexte beziehen. Daneben finden sich detaillierte analytische Skalen ebenso wie Referenzrahmen zu Lehrplaninhalten und Bewertungskriterien für Lernstadien (z.B. die EUROCENTRES-Skala oder die *British National Language Standards*). Viele der Quellskalen stammen aus dem Kontext des *Foreign Service Institute* (FSI), aber auch Fulchers empirisch hergeleitete Flüssigkeitsskala wurde aufgenommen (Fulcher 1996) und spielt eine nicht geringe Rolle für die Skalierung. Die Herkunftsskalen sind sehr heterogen (Hulstijn et al. 2010: 14).

In einem nächsten, qualitativen Schritt wurde mit verschiedenen Verfahren eruiert, welche Kategorien Lehrende für die Beurteilung der L2-Kompetenz als relevant und nützlich empfinden, um zu prüfen, ob diese durch die vorhandenen Deskriptoren repräsentiert würden. Es fanden 32 Workshops statt, in denen Lehrende die Deskriptoren sowohl nach Schwierigkeit als auch nach Kategorie sortierten und sie hinsichtlich ihrer Eignung zur Bewertung beurteilten (vgl. Vogt 2011).

Anschließend wurden die konsistentesten Deskriptoren zur Erstellung von miteinander verankerten Fragebögen verwendet, mit deren Hilfe die teilnehmenden Lehrpersonen eine große Anzahl an Lernern einschätzen sollte. In diesen Fragebögen besaß jeder Deskriptor eine eigene Skala. Außerdem wurden mit Hilfe von Videodaten weitere Lernerleistungen beurteilt.

Darauf aufbauend konnte das quantitative Verfahren der Multi-Facetten-Rasch-Analyse angewandt werden, um eine einzige Gesamtskala zu erstellen. Eine solche Rasch-Analyse ist ein probabilistisches statistisches Verfahren aus der Item Response-Theorie, das sich im Sprachtesten in verschiedenen Bereichen zunehmender Beliebtheit erfreut. Mit seiner Hilfe ist es hier möglich gewesen, Personen (genauer: ihre fremdsprachliche Kompetenz), 'Items' (d.h. die Deskriptoren) und deren Schwierigkeit auf derselben Skala anzuordnen; all diese so genannten Facetten werden dabei als Aspekte einer Testsituation betrachtet. Eine Hauptvoraussetzung für die Anwendbarkeit der Rasch-Analyse ist die zugrunde liegende Unidimensionalität, d.h. die beteiligten Items müssen nach denselben Messprinzipien funktionieren, selbst wenn sie komplexe Konstrukte repräsentieren (vgl. Fulcher 2003: 108ff.). Dieser Schritt wurde in einer zweiten Projektphase (mit anderen LehrerInnen, Sprachen und Fertigkeiten) repliziert; sehr hohe Korrelationswerte konnten dabei erreicht werden (North 2000: 339).

Das Ergebnis des Skalierungsverfahrens ist eine Skala der kommunikativen Sprachkompetenz, die auf einer Messtheorie beruht und deren Stufen gleich

weit voneinander entfernt sind (Fulcher 2003: 111). Die zehnstufige Kompetenzskala enthält skalierte Beschreibungen zu mündlicher und schriftlicher Interaktion, zur mündlichen Produktion, zum Hörverstehen sowie eine Skala zum Leseverstehen (Schneider/North 2000: 99). Aus den zehn Niveaus des *Schweizer Projekts* wurden im GeRS sechs, eine Setzung, bei der auf politische Hintergründe Rücksicht genommen worden sei, so Schneider/North (2000:153).

Einige Aspekte der Kompetenz erwiesen sich als nicht oder nur schwierig skalierbar (North 2000: 318); dazu gehört die soziokulturelle Kompetenz. Die Skala zur soziolinguistischen Angemessenheit wurde in einem Folgeprojekt (1999–2000) entwickelt und dem GeRS hinzugefügt.

Das Erstellungsverfahren ist als '*measurement-driven*' einzuschätzen (vgl. Fulcher et al. 2011: 7, vgl. Kapitel 1.1.2); Lehrer-Entscheidungen werden als Daten behandelt, während das Rasch-Modell als Schiedsrichter fungiert, indem es die aufzunehmenden und auszuschließenden Deskriptoren bestimmt. Die *horizontale Dimension*, also die verschiedenen Aspekte der Sprachkompetenz (Kapitel 5) und die kommunikativen Aktivitäten (Kapitel 4) geht zwar auch auf theoretische Reflexionen zurück (North 2000), beruht aber letztendlich auf Kategorisierungen der beteiligten Lehrpersonen und ist somit nicht eigentlich theoriebasiert. Auch mussten wie erwähnt einzelne Aspekte der kommunikativen Kompetenz bei der Skalierung aus methodisch-statistischen Erwägungen außen vor gelassen werden, während andere Deskriptoren erst während der Kalibrierung neu geschrieben wurden (vgl. dazu kritisch Vogt 2011). Die *vertikale Dimension* der GeRS-Gesamtskala ist ebenfalls auf Bewertereindrücke zurückführbar; die am konsistentesten eingesetzten Deskriptoren werden als qualitativ am hochwertigsten betrachtet. Weder bei der Herleitung der horizontalen Kategorien noch deren vertikalen Abstufungen wurden empirische Lernerdaten analysiert, während theoretische Überlegungen nachrangig waren: die Skalierung der GeRS-Deskriptoren folgte dem *Leitprinzip der Praktikabilität*.

1.2.2 Analyse der GeRS-Skalen

Im Folgenden sollen die GeRS-Skalen anhand der in Kapitel 1.1.1 vorgestellten Kategorien beschrieben und kritisch diskutiert werden. Dabei wird sich zeigen, dass eine eindeutige Zuordnung nicht in jedem Falle möglich ist. Der Fokus wird auf den Skalen des fünften Kapitels liegen, die auch Gegenstand der weiteren Analysen dieser Arbeit sind.

1.2.2.1 Generalisierbarkeit: Eine Kluft im Referenzrahmen

Beanspruchen die GeRS-Skalen, ein Bild der zugrunde liegenden Kompetenz (*ability/proficiency/competence*) zu geben (*ability-based approach*), oder versuchen sie, die verschieden ausgeprägte, möglichst lebensnahe Bewältigung von Sprachaufgaben zu beschreiben (*real-life* bzw. *can-do*)? In der Regel werden die Skalen letzterem Ansatz zugeordnet (Bachman 2011; Fulcher 2003: 107); dies legt ja auch bereits die Form der ‚Kann-Beschreibungen' der Deskriptoren nahe. Auch Brian North schreibt, dass

> "(…) a common framework scale needs a "real-life" dimension" (North 2000: 19)

und damit einen Bezug auf relevante kommunikative Aktivitäten. Gleichzeitig sei jedoch zur angemessenen Beschreibung qualitativer Aspekte des Sprachgebrauchs eine '*ability*'-Dimension nötig (ders.: 20, 28). Die GeRS-Skalen sind, anders als bei ihrer Interpretation allgemein betont, *nicht ausschließlich* 'real-life'- bzw. '*can do*'*-basiert, sondern* haben durchaus *auch* den Anspruch, aus '*ability*'*-Perspektive* Aussagen über Kompetenzen von Sprechern zu ermöglichen. Diese zweifache Ausrichtung wird in der Gliederung des GeRS auch deutlich gespiegelt. Während die Skalen des vierten Kapitels konkrete kommunikative Aktivitäten erfassen ('real life'), beziehen sich die Skalen des fünften Kapitels auf Aspekte der kommunikativen Sprachkompetenz ('ability'). Die Form der Kannbeschreibungen ändert nichts daran, dass die Skalen in Kapitel 5 den Anspruch erheben, Aussagen zur Kompetenz zu ermöglichen; sie beziehen sich weder auf 'TLU domains' (vgl. Bachman/Palmer 1996) noch auf konkrete kommunikative Aktivitäten.[32]

Es ist nahe liegend, dass die Erarbeitung von Kategorien zur Skalierung in Kapitel vier und fünf in zwei Stufen erfolgen musste:

> "The approach taken in this study and in the Council of Europe Common European Framework is to separate the consideration of Categories for Competence/Proficiency from Categories for Communicative Activities, with Strategy Use seen as a hinge linking the two" (North 2000: 53).

---

32 Der GeRS formuliert sehr knapp über das Verhältnis der Skalen aus Kapitel 5 zu den dort beschriebenen kommunikativen Kompetenzen. Bei den linguistischen Kompetenzen sind die Autoren besonders vorsichtig, es findet sich lediglich die vage Formulierung „Fortschritte, die Lernende bei der Nutzung sprachlicher Mittel machen, lassen sich in Skalen fassen" (EUROPARAT 2001b: 110). Bei den soziolinguistischen und pragmatischen Kompetenzen hingegen ist eindeutiger formuliert, dass es eben diese sind, die in den Skalen operationalisiert wurden (EUROPARAT 2001b: 121 und 123).

In der vorliegenden Arbeit liegt der Fokus ausschließlich auf Skalen des fünften Kapitels, weswegen im Folgenden die Kategorienbildung für die Skalen zu kommunikativen Aktivitäten nicht diskutiert wird (vgl. North 2000: 98–115). An dieser Stelle sei jedoch darauf verwiesen, dass innerhalb des Referenzrahmens eine methodische und inhaltliche Kluft verläuft – Skalierungen im 'real-life' – und im 'ability-approach' erfordern nicht nur grundlegend verschiedene Erstellungsmethoden, sondern die resultierenden Skalen unterscheiden sich auch hinsichtlich der meisten der hier diskutierten Charakteristika erheblich. Diese Trennung zwischen Kapitel 4 und Kapitel 5 wird aber Lesern des Referenzrahmens nicht bewusst gemacht; nur über Umwege können sie an Informationen gelangen. Außerdem wurden zwar die Kategorien der Kapitel 4-Skalen getrennt von denen der Kompetenzskalen in Kapitel 5 hergeleitet; das anschließende Kalibrierungsverfahren erfolgte jedoch für alle Deskriptoren *gleichzeitig*. Es ist fraglich, inwiefern ein einziges zusammenfassendes, wenn auch komplexes, Vorgehen wirklich beiden Ansprüchen gerecht werden kann.[33]

1.2.2.2 Kontextfreiheit und Kontextgebundenheit

Während Skalen im *can do*-Ansatz kontextualisiert sein müssen, sollten *ability*-Skalen kontextfrei sein. Als Konsequenz des zweifachen Anspruchs der GeRS-Skalen (Kapitel 1.2.2.1) muss demzufolge auch mit dem Kontext gänzlich unterschiedlich verfahren werden. North bringt die Anforderungen auf den Punkt:

> "A framework scale ideally needs to be context-free in order to accommodate generalizable results from different specific contexts, yet at the same time the descriptors on the scale need to be context-relevant, relatable or translatable into each and every relevant context, and appropriate for the functions serving that context. (…) this means that the descriptive scheme of the framework and descriptors need to (…) be relevant to the contexts of the learning population concerned, although these cannot be predicted with any certainty (…)" (North 2007: 658).

33 Ein weiteres Problem liegt in der Tatsache, dass die Deskriptoren (North 2000: 317) als behavioral aufgefasst und dann nicht ohne weiteres verwendet werden könnten, um Aussagen über die kommunikative Sprachkompetenz zu treffen (vgl. Alderson et al. 2004: 2). Dieser Einwand kann teilweise durch NORTHs Konkretisierung entkräftet werden, wonach Skalen, die das Verhalten von Lernern beschreiben, nicht zwingend 'behavioral' seien, sondern lediglich eine funktionale Perspektive einnähmen, die durchaus Rückschlüsse, die über bloße Performanz hinausgingen, erlaubten (NORTH 2000: 25–28).

Dieser Spagat gelingt indes nur sehr bedingt. Claudia Harsch (2005: 213f., 221) charakterisiert die Skalen als tendenziell kontextfrei und kommt sehr kritisch zu dem Schluss, dass sie in der jetzigen Form nur als Bezugspunkt zur Entwicklung neuer Skalen für konkrete Kontexte verwendet werden könnten. Ihrer Auffassung zufolge bedroht der allgemeine Charakter der Deskriptoren die Verwendbarkeit der Skalen in konkreten Kontexten. Dies treffe insbesondere auf die Skalen des vierten Kapitels zu. Die empirischen Analysen dieser Arbeit haben zudem gezeigt, dass auch für die Wortschatz-Skalen des fünften Kapitels erhebliche Probleme entstehen (vgl. Kapitel 3.2.4 und 3.3.5).

1.2.2.3 Funktionen und Zielgruppen der GeRS-Skalen

Die Frage nach den Funktionen und Zielgruppen der GeRS-Skalen hängt eng mit dem gerade dargestellten Problem der Kontextgebundenheit zusammen. Es wurde kritisiert, dass weder im GeRS noch in den zugehörigen Publikationen (wie z.B. North 2000) angegeben wird, zu welchem Zweck die Skalen eingesetzt werden sollen und können (z.B. Harsch 2005: 186):

> "non-purposive (…). It does not detail – and perhaps this was an intent – particular contexts in which it would be used" (Fulcher/Davidson 2007: 232).

North bemerkt dazu recht allgemein und nicht unmissverständlich, die Skalen seien zu nahezu allen Zwecken (Alderson 1991) einsetzbar:

> "Therefore, in order to provide transparent reporting of achievement, a common framework scale should incorporate descriptors not just for those aspects of competence/proficiency which are of interest to insiders (diagnosis-oriented (…)), but also descriptors for task completion. The latter could be used by insiders as a source of content for syllabus organisation and continuous assessment (constructor-oriented (…) as well as, probably in a summarised form, for reporting results to outsiders (user-oriented (…))." (North 2000: 62)

Mehr Informationen finden sich jedoch diesbezüglich bei North nicht. Harsch (2005: 214–217) stellt alle Hinweise zusammen, die sich im GeRS auf mögliche Funktionen der Skalen beziehen: Dort wird etwa betont, die Skalen könnten helfen, Erwartungen an Lernziele zu beschreiben (Europarat 2001b: 27); weitere Funktionen liegen in der Beschreibung von Lernfortschritten (Europarat 2001b: 39) sowie in der Beurteilung von Leistungen (Europarat 2001b: 28) und in der Abstimmung von Lernzielen und -materialien (Europarat 2001b: 28). Zudem böten GeRS-Skalen Hilfe bei der Erstellung von Prüfungen, in der Herstellung von Vergleichbarkeit und in der Hilfe bei der Formulierung von Standards (ebda.). Harsch kritisiert dies und kommt (2005: 207) ebenso wie Fulcher (2004: 264) zu der Ansicht, dass die Skalen

im Referenzrahmen in der vorliegenden Form am ehesten als *reporting scales* für Verwender geeignet sind. Jones/Saville (2009) und Harsch (2005: 217ff.) konstatieren, dass GeRS-Skalen nicht ohne weiteres als Bewertungsinstrumente eingesetzt werden können; der Referenzrahmen selbst gibt hingegen an, man könne die Skalen nutzen, um Beurteilungsraster abzuleiten (Europarat 2001b: 175f.). Harsch äußerst sich äußerst skeptisch gegenüber der Möglichkeit, mit den GeRS-Skalen in Tests und Prüfungen zu arbeiten (2005: 218f.) und vor allem auch gegenüber dem Einsatz von GeRS-Skalen aus dem fünften Kapitel zur Kompetenzbeschreibung. Sie moniert zunächst, dass die Voraussetzung, unabhängige und positive Deskriptoren zu verwenden, nicht durchgängig erfüllt sei (dies.: 220). Dann kritisiert sie:

> „Wie man jedoch eine einzelne Performanz auf Kompetenzniveaus einordnen will, bleibt fraglich, zumal in den Skalenanalysen gezeigt wurde, dass die Kategorien nicht immer stringent in den Skalenniveaus beschrieben sind, die Basis der Beschreibungen nicht transparent ist und es sich dabei um generalisierte Kompetenzbeschreibungen handelt. Diese können (…) nicht genutzt werden, um konkrete Performanzen zu bewerten; dazu müssten die Skalen schon konkrete Performanzmerkmale beschreiben" (ebda.).

Die Skalen des Kapitels 5 des GeRS entsprechen nicht Harschs Definition von (beurteilerorientierten) *rating scales*. Diese müssten abhängig von der jeweiligen Aufgabe Performanzmerkmale enthalten (dies.: 157).

Harschs vielleicht etwas überspitzte Kritik verweist auf ein grundsätzliches Dilemma, das nicht ausschließlich dem GeRS angelastet werden kann und von diesem kaum gelöst werden kann. Es ist nämlich nicht klar, wie *ability*-basierte Skalen (Kapitel 5 des GeRS) beurteilerorientiert sein können, wenn dazu die Unterbringung von Performanzmerkmalen gehört, für die wiederum es aus der Spracherwerbsforschung nicht ausreichend Belege gibt. Die Forderung nach dem Einbezug möglichst konkreter sprachlicher Oberflächenphänomene, die eine bestimmte Niveaustufe oder einen bestimmten Erwerbsfortschritt anzeigen sollen, ist zwar berechtigt, aber leider bislang theoriebegründet nur lückenhaft möglich. Würde man in einer Skala aus Kapitel 5 andererseits konkrete Performanz *szenarien* (und das heißt eine Bindung an bestimmte Sprachaufgaben) einbauen, würde dies zwar Einschätzungen durch Beurteiler bestimmt erleichtern, jedoch die Generalisierungsmöglichkeiten gleichzeitig erheblich einschränken. Außerdem ist bei *ability*-basierten Skalen anzumerken, dass die Abstraktion einer konkreten Performanz auf eine zugrunde liegende Kompetenz *immer* heikel und spekulativ sein muss, auch wenn eine Vielzahl an Aufgaben als Grundlage dienen kann. Sicherlich, soviel lässt sich zweifellos sagen, erschwert der allgemeine Charakter der GeRS-Skalen insbesondere des 5. Kapitels die direkte Zuordnung von GeRS-Niveaus zu einer Performanz (vgl. Kapitel 3.2.4 und 3.35).

Es ist also davon auszugehen, dass die GeRS-Skalen, sollen sie für konkrete Zwecke, also etwa als Bewertungsinstrumente, eingesetzt werden, *modifiziert* werden müssten. In welcher Art das geschehen soll, wird im Referenzrahmen mit einer Ausnahme nicht angegeben.[34]

Durch diese Vagheit entsteht ein weiteres *Dilemma*: Verfahren, bei denen die GeRS-Skalen verändert werden, sind zwar womöglich dem jeweiligen konkreten Kontext angemessen, laufen aber Gefahr, am Ende Ergebnisse zu produzieren, die kaum noch transparent mit den GeRS-Skalen zusammenzuhängen (vgl. Vogt 2011).

Die Skalen des GeRS sind in der Regel analytisch, wenn sie einzelne Aspekte der sprachlichen Kompetenz oder einzelne kommunikative Aktivitäten zusammenfassen. Die Globalskalen als Synthesen mehrerer analytischer Skalen sind hingegen als holistisch aufzufassen.

1.2.2.4 Kriteriums- und Normorientierung

Die Skalen des GeRS sind prinzipiell kriteriumsorientiert, stellen zudem selbst einen externen Standard da; da aber nicht alle Deskriptoren wirklich unabhängig funktionieren, werden die empirischen Analysen in dieser Arbeit zeigen, dass gelegentlich der Einbezug einer Gruppe notwendig ist, um zu Entscheidungen zu kommen, man also in Abhängigkeit von der Qualität der Deskriptoren von einer **verdeckten Normorientierung** sprechen muss. Sobald gewisse quantifizierende Angaben zu Performanzmerkmalen auftreten wie z.B. ‚wenig lange Pausen', müssen für eine Beurteilung reale Vergleichssprecher hinzugezogen werden. Außerdem sind einige Niveaubeschreibungen (gerade bei den Wortschatzskalen) derart knapp und allgemein gehalten, dass ein Abgleich angrenzender Niveaustufen notwendig ist und Entscheidungen nicht ‚für Niveau x', sondern ‚gegen Niveaus y und z' getroffen werden müssen (vgl. Kapitel 3.2.4).

1.2.2.5 Die Deskriptoren

Im Folgenden soll auf einige Eigenschaften der GeRS-Deskriptoren eingegangen werden. Die Verantwortlichen des *Schweizer Projekts* bezeichnen Deskriptoren

34 Der Referenzrahmen bietet mit Tabelle 3 ein Beispiel für eine beurteilerorientierte Skala (Europarat 2001b: 37f.). Das Raster stellt verschiedene Deskriptoren zusammen und wird zur Beurteilung mündlicher Performanzen mit Bezug auf die Kompetenz empfohlen (vgl. North 2005: 48). Ein ähnliches Bewertungsinstrument wurde in dieser Arbeit verwendet.

nur dann als ‚gut' und ‚skalierbar', wenn das Können positiv formuliert ist, sie für sich allein genommen Sinn ergeben und eine Ja/Nein-Entscheidung ermöglichen. Ferner darf ihre Interpretation nicht abhängig sein von Beschreibungen des gleichen Niveaus, die in anderen Skalen vorkommen, und auch nicht von Deskriptoren angrenzender Niveaustufen. Außerdem wird als vorteilhaft angesehen, wenn Unterschiede der Niveaus nicht nur durch quantifizierende Abstufungen markiert werden, wenig Fachterminologie vorkommt und die Deskriptoren zudem kurz, konkret und klar sind (Schneider/North 2000: 89; vgl. Anhang A in Europarat 2001b). Weitere Skalierungsprinzipien des GeRS lagen darin, nur selten und dann sehr vorsichtig von Fehlern zu sprechen, da diese nur sehr ungenau bzw. sogar ganz unangemessen den Fortschritt des Spracherwerbs beschreiben würden. Auch wurde versucht, keine Deskriptoren aufzunehmen, die fälschlicherweise ein Entwicklungs-Kontinuum von einfacher zu komplexer Sprache (bzw. von unmarkierten zu markierten Formen) suggerieren (North 2000: 83). Die Autoren der Projektpublikation merken an, dass Ergebnisse aus der L2-Forschung zu etwaigen generellen Spracherwerbssequenzen für die Skalierung von Deskriptoren nur eingeschränkt nutzbar seien (Schneider/North 2000: 32). Zudem bemängelt North (2000: 84), dass der Bedeutung von formelhaften Sequenzen in Skalen häufig nicht angemessen oder ausreichend Rechnung getragen werde. Es wurde Wert darauf gelegt, die Kompetenzniveaus nicht durch rein qualitative Beschreibungen zu trennen, da dann keine Kriteriumsorientierung mehr vorliege. Außerdem sei ein solches Verfahren aus messmethodischen Gründen bedenklich:

> "(…) adjacent descriptors which are distinguished solely by word-processing a different qualifier do not meet the need for "incisiveness" or "definiteness" which is a requirement for a valid measurement scale" (ders.: 86).

Die oft kritisierte Tatsache, dass die GeRS-Skalen nicht auf jedem Niveau dieselben sprachlichen Charakteristika beschreiben, steht diesen Anforderungen nicht entgegen und resultiert aus der Skalierungsmethode; die GeRS-Skalen bestehen jeweils aus Deskriptoren, die mehreren verschiedenen ursprünglichen Quellen entstammen, so dass die Fragmentierung der GeRS-Skalen vorprogrammiert war.

Im Laufe der letzten Jahre sind Anstrengungen unternommen worden, um fehlende Deskriptoren zu ergänzen, während einzelne Kategorien der kommunikativen Kompetenz bislang noch gar nicht haben beschrieben werden können (Sprachmittlung; soziokulturelle Kompetenz). Die statistische Qualität der einzelnen Deskriptoren ist nachvollziehbar (Schneider/North 2000: 235ff.), wenn auch nicht im Referenzrahmen selbst.

Allerdings entsprechen durchaus nicht alle GeRS-Deskriptoren diesen Ansprüchen. Einige Probleme sollen anhand der hier analysierten Skalen zum Wortschatz und zur Flüssigkeit angerissen werden. Problematisch ist eine gewisse Unentschiedenheit bei Deskriptoren des fünften Kapitels, die Aussagen zur Kompetenz ermöglichen sollen: ‚Beherrscht einen begrenzten Wortschatz in Zusammenhang mit konkreten Alltagsbedürfnissen' (Niveau A2, Skala zur Wortschatzbeherrschung). Dieser Deskriptor ermöglicht nur dann eine Ja/Nein-Entscheidung, wenn die Aufgabe, aufgrund derer das 'ratable sample' (Bachman/Palmer 2010) produziert wurde, auch erforderte, konkrete Alltagsbedürfnisse zu erledigen. Auf mehreren Niveaustufen der Skala ist von solchen verschiedenen Sprachfunktionen die Rede, so dass sich nur durch ein mehrstufiges Aufgabengeflecht eine Niveauzuordnung vornehmen ließe. *Die Deskriptoren sind damit nicht allgemein genug für eine ability-Skala, aber auch nicht konkret genug, um real-life-Anforderungen zu entsprechen.* Ein Beispiel für extrem unterschiedliche Konkretisierungen innerhalb ein und derselben Skala zeigt sich in der ohnehin problematischen Skala zur soziolinguistischen Angemessenheit: ‚Kann Filmen folgen, in denen viel saloppe Umgangssprache oder Gruppensprache und viel idiomatischer Sprachgebrauch vorkommt.' (Niveau C1). Diese sehr konkrete Beschreibung steht direkt neben einer überaus allgemeinen (B2+): ‚Kann sich in formellem und informellem Stil überzeugend, klar und höflich ausdrücken, wie es für die jeweilige Situation und die betreffenden Personen angemessen ist.' Zudem bezieht sich die erstgenannte Beschreibung auf rezeptive, die zweite auf produktive Fähigkeiten.[35]

Soll eine Produktion (wie in Tabelle 3 des GeRS vorgeschlagen, vgl. Kapitel 1.2.2.3 und 1.3.4.1) direkt einem GeRS-Niveau zugeordnet werden, entstehen Probleme, wenn die in der Skala vorkommenden Szenarien nicht auftreten, zu allgemein ausgeprägt sind (bspw. machen alle Sprecher viele Pausen oder verwenden alle Sprecher fast ausschließlich Grundwortschatz), oder tatsächliche Deskriptor-Ausprägungen sich innerhalb eines Niveaus widersprechen. In solchen Fällen bedroht die Empirie die Praktikabilität, denn

35 North (1994: 38) ist sich der Gefahr, die entsteht, wenn Deskriptoren aus ursprünglichen Zusammenhängen gerissen werden, durchaus bewusst: "Basically, anyone developing a scale seems to start by seeing what already exists; even the writers of the FSI scale, the "original" scale, did this. As discussed, this raises the problem of descriptors taken out of context and used for a purpose or group for whom they were not intended, and of decisions about what to put where being based on convention."

Bewerter müssen auf idiosynkratische Konzepte oder andere Niveaustufenbeschreibungen rekurrieren.

Ein weiterer Kritikpunkt liegt in der vorwiegend *mündlichen Ausrichtung* der Deskriptoren: bei der Orthographie beispielsweise geht es auf den unteren Niveaustufen um das Buchstabieren, in der Skala zum Wortschatzspektrum ist auf Niveau B2 von ‚Zögern und Umschreibungen' die Rede. Dies schränkt die Anwendbarkeit der Skalen ein. Viele Deskriptoren sind *subjektiv* (z.B. Flüssigkeitsskala, Niveau B2 ‚normales Gespräch mit Muttersprachler ist ohne Belastung einer der beiden Seiten' möglich, Niveau C2, B1+, A2: ‚mühelos'), *selbstreferentiell* (z.B. Flüssigkeitsskala, Niveau C1, ‚Kann sich beinahe mühelos spontan und fließend ausdrücken'), oder *vage* (Wortschatzbeherrschung, C1: ‚kleinere Schnitzer'). Als Resultat der getrennten Erstellung von Skalen und Text des Referenzrahmens *fehlen viele Definitionen im Text*; die Skalen werden teilweise nicht in das Dokument eingebettet, so etwa bei der Wortschatzbeherrschung und bei der Flüssigkeit. Ein weiteres Problem liegt in den teilweise *fehlerhaften Übersetzungen*, die die Übertragbarkeit der Skalen über mehrere Sprachen bedrohen. Beispiele finden sich in den hier betrachteten Skalen, wenn etwa im Englischen A2-Deskriptor der Flüssigkeit von ‚phrases' die Rede ist, angemessen ins Italienische übertragen mit ‚espressioni', im Deutschen aber von ‚Redewendungen' gesprochen wird (Kapitel 2.1.4.2.1).

Als Konsequenz der Skalierungsmethode sind einige Skalen und Deskriptoren von zahlreichen *Ursprungsskalen* abgeleitet (insbesondere die Flüssigkeitsskala), während andere nur auf sehr wenigen Bezugsskalen basieren (etwa die Wortschatzskalen). Diese Unterschiede werden Nutzern nicht verdeutlicht. Es ist über Umwege möglich, die Ursprungsskalen der einzelnen Deskriptoren nachzuvollziehen (vgl. Anhang A). Da die Deskriptoren teils stark verändert wurden, ist aber eine Rückführung auf die ursprünglichen Skalen nicht immer machbar. Auch zeigt sich teilweise eine große Beeinflussung einzelner Skalen durch bestimmte Ursprungsskalen, etwa der FSI-Gruppe (vgl. Kapitel 2.1.4.2.2). Auch ein erheblicher Einfluss von *Eurocentres* ist zu konstatieren:

> "One third of the surviving 212 items come purely from the current writer and Eurocentres, two thirds come either purely or partly from the current writer and Eurocentres" (North 2000: 317).

Ein weiterer wichtiger Aspekt, den Claudia Harsch anspricht, ist die Frage danach, wie *typisch* die in Deskriptoren erwähnten Charakteristika für eine Niveaustufe wirklich sind – beschreiben sie die prototypische Niveaumitte oder stellen

sie eine Grenzziehung dar (Harsch 2005: 149)? Diese Fragen sind für Benutzer von Skalen von zentraler Bedeutung.

1.2.3 Das Validitätsdefizit der GeRS-Skalen

Bereits im vorangegangenen Kapitel hat sich gezeigt, dass die GeRS-Skalen in vielfacher Hinsicht nicht widerspruchsfrei sind bzw. sich nicht leicht hinsichtlich der üblichen Beschreibungskategorien einordnen lassen. Es ist jedoch der im Folgenden entwickelte, als besonders gravierend und grundlegend aufgefasste Kritikpunkt der mangelnden Validität der GeRS-Skalen, der Anlass zur Entstehung dieser Arbeit gegeben hat.[36]

Validität bei Sprachtests kann durch eine gelungene Argumentation für die Angemessenheit von Testinterpretationen hergestellt werden (Messick 1989; vgl. auch Bachman/Palmer 2010; Kane 2001, 2013).[37] Im Gegensatz zu früheren Ansätzen, in denen u.a. zwischen ‚Konstruktvalidität', ‚Inhaltsvalidität' und ‚Augenscheinvalidität' unterschieden wurde, werden diese heute als Facetten einer Gesamtvalidität aufgefasst (vgl. für einen historischen Überblick Kane 2001 oder Chapelle 1999; siehe auch Bachman 2005; Messick 1989). Validität ist

36 In dieser Arbeit wird nicht beabsichtigt, eine möglichst umfassende Kritik des GeRS zu liefern. Deshalb wird auf einige Aspekte, wie etwa die Schwierigkeiten, Tests auf den GeRS zu beziehen (vgl. dazu Weir 2005), nicht eingegangen.

37 Es sind nach Kane (2001: 332ff.) zwei Typen solcher Interpretationen von Testergebnissen möglich. Zum einen ist denkbar, dass Aussagen über beobachtbare Eigenschaften getroffen werden sollen; man würde dann ohne viel Spekulation behaupten, dass eine einzelne Performanz typisch für eine Fertigkeit ('skill') ist, die aus einem Universum möglicher Performanzen besteht. Während Testkonstruktionen in diesem Paradigma aufwändig sein können, ist doch die Anwendung einfacherer Validierungsstrategien möglich, weil eine beobachtbare Eigenschaft interpretiert werden kann, ohne dass (zur Debatte stehende) Theorien mit einbezogen würden (ders.: 333). Zu diesem Paradigma passen tendenziell Skalen mit einem 'real-life'-Fokus (vgl. Kapitel 1.1.1). Zweitens ist möglich, dass Testergebnisse Aussagen über theoretische Konstrukte (z.B. Aspekte der Sprachkompetenz) liefern sollen. Dann muss zunächst mindestens eine Performanz bewertet werden, um einen beobachteten 'Score' zu erreichen (der unabhängig von konstruktirrelevanter Varianz sein muss). Dieser Score entspricht nicht dem Konstrukt selbst, sondern bildet für dieses einen Index (ebda.). Bei solchen Testergebnisinterpretationen ist eine deutlich höhere Anzahl an Inferenzen nötig. 'Ability'-basierte Skalen müssen prinzipiell auf Theorien und Modelle rückführbar sein.

> "(…) an integrated evaluative judgment of the degree to which empirical evidences and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick 1989: 13).

Validitätsaspekte von Skalen werden, wie gesagt, selten explizit fokussiert (vgl. Kapitel 1.1.3). Dies ist wenig erstaunlich, da die meisten Skalen in konkreten Testkontexten entstehen, die dann insgesamt validiert werden:

> "Validity can only be seen as relative to function and context: "What is this test/scale valid for?" rather than "Is this test/scale valid?" (Henning 1990:379, zitiert nach Europarat 1994a: 12).

Die Tatsache, dass die GeRS-Skalen für konkrete Testkontexte erst angepasst werden müssen, entbindet jedoch nicht von der Pflicht, bereits im Vorhinein Nachweise für die *Möglichkeit* ihres validen Einsatzes zu bringen. Wie unten (vgl. Kapitel 1.3.2) weiter ausgeführt wird, werden als grundsätzliche, unabhängig vom Testkontext gültige zentrale Validitäts *voraussetzungen* hier an vorderster Stelle die Rückführbarkeit von Skaleninhalten auf Modelle und Theorien und ihr Bezug auf empirische, authentische Lernersprache verstanden. Eine Rolle spielt auch die Handhabbarkeit von Skalen durch Beurteilende.

Bislang gibt es keine Studie, in der die mangelnde theoretische (mit Ausnahme von Harsch 2005), vor allem aber empirische Validität einzelner GeRS-Skalen untersucht würde. Bestehende Initiativen, beispielsweise zur Illustration der GeRS-Niveaus oder zur Herstellung eines Zusammenhangs von GeRS-Niveaus und Spracherwerb, setzen – anders als die vorliegende Arbeit – an einem Punkt an, an dem die Existenz dieser Niveaustufen nicht mehr in Frage gestellt, sondern als gegeben hingenommen wird.

Die sich anschließenden kritischen Ausführungen zu theoretischen (Kapitel 1.2.3.1) und empirischen (1.2.3.2) Aspekten der GeRS-Skalen beziehen sich auf die kompetenzbezogenen Skalen des fünften Kapitels des Referenzrahmens, wobei die in dieser Arbeit näher untersuchten Skalen zum Wortschatz und zur Flüssigkeit vorrangig betrachtet werden. Es sollen aber einige explizit wertschätzende Bemerkungen zum GeRS und seinen Skalen vorausgeschickt werden (vgl. Kapitel 1): Bei aller berechtigten Kritik muss nämlich anerkannt werden, dass ein derart transparent erstelltes Skalensystem im internationalen Vergleich die absolute Ausnahme darstellt; im Allgemeinen erhalten Nutzer überhaupt keine oder nur sehr wenige Informationen über die Art und Weise der Erstellung von Skalen, aufgrund derer für sie teils sehr wichtige Entscheidungen getroffen werden (vgl. Brindley 1998; Fulcher 2003; Knoch 2011; McNamara 1996; Upshur/Turner 1995).

Außerdem ist die Gründlichkeit der Skalierungsmethode beispiellos, wie auch kritische Autoren zugestehen:

> "It is fair to say that the resultant scales are probably the best researched scales of foreign language in the world, although perhaps not (yet) the most widely used – that award probably goes to one of the source scales, the FSI, ILR, ACTFL family of scales" (Hulstijn et al. 2010: 15).

Auch trifft die Einschätzung von Brian North sicherlich zu, der den GeRS-Skalen das Potenzial zusprach,

> "(…) to exert a positive influence on the orientation, organisation and reporting of language learning" (North 2000: 13).

Seit Erscheinen des GeRS sind sehr viele Anstrengungen zur Verbesserung und Standardisierung von Prüfungen, Sprachunterricht usw. unternommen worden, die sich auf den Referenzrahmen zurückführen lassen. Auch sind viele Portfolio-Projekte entstanden, die mit den Skalenniveaus arbeiten, während gleichzeitig zahlreiche Europarat-Publikationen die Qualität von auf den GeRS zu beziehenden Tests erhöhen helfen sollen oder in Fallstudien das Funktionieren des GeRS in konkreten Kontexten beschreiben.[38] All diese Anstrengungen dienen dazu, den Referenzrahmen mit seinen Skalen zu verbreiten und zu etablieren, während Schwächen durchaus anerkannt und teils konkrete Lösungsvorschläge angeboten werden.

1.2.3.1 Zur theoretischen Verankerung der GeRS-Skalen (Kapitel 5)

Der Anspruch, Aussagen über die Sprachkompetenz zu treffen, zieht die Notwendigkeit des Bezugs der GeRS-Skalen auf eine Theorie des L2-Erwerbs bzw. auf ein Modell der kommunikativen L2-Kompetenz nach sich (vgl. Kapitel 1.2.2.1 und 1.2.). Dies wird grundsätzlich durch das Fehlen umfassender theoretischer Lösungsvorschläge in diesem Bereich erschwert (North 2000: 20; vgl. Kapitel 1.1.2). Im *Schweizer Projekt* wurde dennoch versucht, vor der Skalierung der Deskriptoren einen theoretischen Bezugsrahmen abzustecken. North (1997, 2000) spricht eine Reihe von Modellierungen der kommunikativen Sprachkompetenz an. Es wird letztlich aber nicht klar, auf welchen Begriff bzw. welches Modell die Skalierung genau zurückgeht:

> "(…) the approach adopted takes the more behavioural view of proficiency outlined by Parks, broadening the definition of Pragmatic Competence to include Skehan's

38 Vgl. die Seiten des Europarats, http://www.coe.int/t/dg4/linguistic/dnr_en.asp, Oktober 2013.

Ability for Use, Spolsky's Knowing how to use a language, and Fillmore's Fluency (…)" (North 2000: 54).

Die Kategorien der L2-Kompetenz, zu denen die zu skalierenden GeRS-Deskriptoren geordnet werden sollen, werden dann aus den relativ ähnlichen Modellen von Canale und Swain (1980), Van Ek (1986, Van Ek/Trim 1990) und Bachman (1990) (bzw. Bachman/Palmer 1996) übernommen (vgl. North 2000: 74ff.): dabei handelt es sich um die strategische, linguistische, pragmatische und soziokulturelle Kompetenz. Gerade die Modelle von Canale (1983) (bzw. Canale/Swain 1980) sowie von Bachman/Palmer 1996 (bzw. 2010; Bachman 1990) sind sehr einflussreich:

Das deskriptive Modell kommunikativer Kompetenz von *Canale (1983) bzw. Canale/Swain (1980)* besteht aus zwei Bestandteilen. Die kommunikative Kompetenz (die Autoren lehnen sich hier an Hymes (1972 an, vgl. Shohamy 1996: 142) zerfällt in die grammatische, die strategische und die soziolinguistische Kompetenz, die eine Wissens- und eine Fertigkeiten-Seite haben. Kompetenz wird als Wissen definiert; der '*ability*'-Aspekt wird bewusst außen vor gelassen (Schneider/North 2000: 25). Strategien sind zunächst relativ knapp und tendenziell defizitorientiert bestimmt als Verfahren, mit denen man versucht, Kommunikationsprobleme zu lösen (vgl. Fulcher/Davidson 2007: 38). Die zweite, 1983 weiter ausgearbeitete Modellkomponente ist die 'actual communication' (Canale/Swain 1983: 5), wobei es sich um ein Performanz-Modul handelt. Nun tritt auch die Diskurskompetenz als neue Kategorie hinzu (Canale/Swain 1983: 11).

Eine weitere wichtige Säule der Skalierung ist das prozessorientierte Modell der *Communicative Language Ability*[39] (CLA) von Bachman/Palmer (1996).[40] Bachman (1990) bzw. Bachman/Palmer (1996, 2010) beziehen die verschiedenen Komponenten der CLA funktional einerseits aufeinander, andererseits stets auf ihre Interaktion mit der Test-Aufgabe (Bachman/Palmer 1996: 62). 'Knowledge' und 'skill' sind klar definiert (Fulcher/Davidson 2007: 42).[41] Die vier Fertigkeiten (*language skills*) gehören bei Bachman/Palmer nicht zur Sprachkompetenz

39 Im Deutschen finden sich die Übertragungen ‚kommunikative Sprachfähigkeit' und ‚kommunikative Kompetenz' (Glaboniat 1998).

40 Eine der wenigen kritischen Stimmen zum Modell von Bachman/Palmer 1996 (bzw. neu 2010) ist Mcnamara 2003.

41 Die Übersetzung dieser Begriffe ist nicht unproblematisch. 'Skill' kann als ‚Fertigkeit, Fähigkeit' übersetzt werden, so aber auch 'ability'. An Stellen, wo eine Übersetzung dieses oder anderer Begriffe zu Missverständnissen führen könnte, wird der Originalbegriff mit angegeben.

(*language ability*), sondern werden aufgefasst als *aufgaben*bezogene Realisierung dieser Kompetenz im Kontext (Bachman/Palmer 1996: 75f.). *Language Ability* ist bei Bachman/Palmer eine individuelle Eigenschaft. Für das Testen sind weitere individuelle Charakteristika zu berücksichtigen, nämlich persönliche Eigenschaften, Sachwissen und affektive Schemata (dies.: 78). Die kommunikative Sprachkompetenz besteht aus zwei Modulen (ebda.). Zum einen gibt es den aus organisatorischem und pragmatischem Wissen[42] zusammengesetzten Bereich des *Sprachwissens*. Organisatorisches Wissen beinhaltet ein formales (grammatisches) und ein eigentlich pragmatisches Element, das textuelle Wissen (*textual knowledge*), bei dem es um Kohärenz und Kohäsion bzw. den Bereich der rhetorischen Organisation geht. Das pragmatische Wissen hat eine funktionale und eine soziolinguistische Komponente. Die Kenntnis sprachlicher Funktionen ist das Bindeglied zu Zielen der Kommunikation. Das Wissen um soziolinguistische Aspekte betrifft v.a. die diaphasische Angemessenheit der Sprache, aber auch das Wissen um diatopische Varietäten und bildhaftes Sprechen (dies.: 65–68 und 70ff.). Das zweite, ebenfalls zentrale Modul ist die *strategische Kompetenz* (Fulcher/Davidson 2007: 44). Bachman/Palmers Modell ist sehr umfassend, denn es wird nicht nur die kommunikative Sprachkompetenz selbst dargestellt, sondern eine Reihe weiterer Faktoren wie das Sachwissen (*topical knowledge*), persönliche Eigenschaften der Testperson, Besonderheiten der Aufgabe und des *Setting* (Bachman/Palmer 1996: 63). Die Kategorien des Modells beruhen teilweise auf empirischen Untersuchungen bzw. sind durch solche bestätigt worden (Schneider/North 2000:28).

Im Folgenden soll die theoretische Verankerung der GeRS-Skalen kritisch diskutiert werden. Unter einem zufrieden stellenden, wenn auch nicht erschöpfenden Theoriebezug wird hier verstanden, dass Skalen Erkenntnissen der Forschung zumindest nicht widersprechen dürfen, dass zentrale Elemente und Ergebnisse der Forschung aufgegriffen werden, und dass den Skalen ein eindeutig identifizierbares Konstrukt zugrunde liegen sollte; es ist also aus dieser Warte für eine Skalierung nicht nötig, über eine ‚fertige' Erwerbs- oder Kompetenztheorie zu verfügen.

Zweifelsohne haben die Autoren des *Schweizer Projekts* in einem sehr umfassenden Ansatz zahlreiche dem Erkenntnisstand der Forschung entsprechende theoretische Bezüge berücksichtigt. Vor die immense Herausforderung gestellt, ein allgemeines Referenzwerk an stets unvollständige Forschungsergebnisse koppeln zu müssen (vgl. Jones/Saville 2009; Kane 2001), wurden deutliche

42 Zwischen Bachman/Palmer 1996 und Bachman 1990 gibt es einige Unterschiede in der Terminologie; hier wird dem neueren Werk gefolgt.

Anstrengungen unternommen, theoretisch begründet vorzugehen. Trotzdem lässt sich aus mehreren Gründen, die mit dem Stand der Forschung – von der kaum zu hoffen (oder zu befürchten) steht, dass sie jemals ‚abgeschlossen' sein könnte – gar nichts zu tun haben, insgesamt nicht davon ausgehen, dass ein zufrieden stellender Theoriebezug der Skalen vorliegt:

(1) Die theoretisch gestützten Kategorien werden im Laufe des Projekts auf bestehende Deskriptoren angewandt; d.h. Deskriptoren, die aus verschiedensten Skalen stammen, wurden *ex post* Kategorien der Kompetenz zugeordnet, mit denen sie ursprünglich womöglich nichts zu tun hatten. Der Theoriebezug ist also, wenn die Deskriptoren selbst nicht theoriegestützt begründet sind, als *mittelbar* zu begreifen. Es müsste jeweils klar sein, inwiefern die Inhalte der Deskriptoren in Übereinstimmungen mit Erkenntnissen der Forschung zu bringen sind. Eine solche Überprüfung fand aber nicht statt. Den verwendeten Deskriptoren wurde sehr großes Vertrauen entgegengebracht.

(2) Es kann nicht davon ausgegangen werden, dass den Skalen ein eindeutiges *Konstrukt* zugrunde liegt, nur weil es Praktikern möglich ist, die verschiedenen Deskriptoren bestimmten Kompetenzkategorien zuzuordnen:

> „Es entstand eine Fassung, die den Konsens europäischer Fremdsprachenlehrender spiegeln könnte – aber keine Fassung, die einer linguistischen oder lerntheoretischen Analyse standhielte. Folglich kann man auch nichts dagegen einwenden, wenn die Benutzer des GeR eher mit „gefühlten" Skalen arbeiten (….). Wo ein geeigneter Deskriptor fehlt, kann man sich schnell einen neuen basteln." (Quetz 2007: 49).

(3) Drittens beschränken sich die Kategorien auf das, was ohnehin verfügbar war - es ist weder klar, ob die kommunikative Kompetenz in dieser Weise *umfassend* umschrieben wird, noch, wie *bedeutsam* die verwendeten Kategorien für die Kompetenz tatsächlich sind. Schließlich mussten aus praktischen bzw. methodischen Erwägungen einige zentrale Bereiche der kommunikativen L2-Kompetenz ohne Beispielskalen bleiben.

(4) Außerdem enthalten die *Modelle kommunikativer Kompetenz***,** die den Skalen als Referenz dienen, *nur sehr knappe* oder – wie im Fall der Flüssigkeit – gar keine genaueren *Beschreibungen* der Kategorien.[43] Sie sind zwar zur Kategorienbildung nützlich (die im Übrigen Gefahr läuft, in eine funktional unterspezifizierte

43 Bachman/Palmer (1996) und Canale/Swain (1980) erwähnen die Flüssigkeit nicht, während ihr eine prominente Stellung unter den Skalen zukommt, weil Bewerter mit ihr besonders gut umgehen konnten. Der Bezug der GeRS-Skalen zu den Modellen kommunikativer Kompetenz ist an dieser Stelle fraglich. Offenbar wurde praktischen Erwägungen hier größere Bedeutung eingeräumt.

Taxonomie zu münden); es ist aber fraglich, ob die Modelle grundsätzlich für eine *ability*-basierte Skalierung hinreichend spezifisch sind. Was es nämlich genau bedeutet, flüssig zu sprechen, über einen großen oder kleinen Wortschatz zu verfügen oder Höflichkeitsfloskeln zu beherrschen, ist weniger Bestandteil von Modellen der L2-Kompetenz, dafür aber Gegenstand zahlreicher Studien der Angewandten Linguistik bzw. der Fremdsprachenerwerbsforschung. Im Kontext dieser Arbeit sind insbesondere die *linguistische und die pragmatische Komponente* der kommunikativen Kompetenz von Bedeutung.[44] Bezüglich beider Kategorien, insbesondere aber bei der linguistischen Kompetenz, finden sich in den Projektpublikationen (North 2000; Schneider/North 2000) weniger theoretische Erläuterungen als Ausführungen zu möglichen Schwierigkeiten bei der Formulierung von Deskriptoren (vgl. Kapitel 1.2.2.5 und North 2000: 80ff.). Die Gelegenheit, die selbst in den genannten Bezugsmodellen tendenziell unterspezifizierten Aspekte der linguistischen und pragmatischen kommunikativen Kompetenz näher zu beschreiben, wurde nicht genutzt. Soll der Modellbezug nicht nur eine *Alibi-Funktion* erfüllen, hätte – trotz des immensen Mehraufwands – bei der Skalierung zu jedem einzelnen der zu skalierenden Bereiche eine intensive Lektüre erfolgen müssen, um den aktuellen Forschungsstand zu erfassen (auch in Ermangelung von Modellen oder Theorien).[45]

(5) Es ist jenseits des Bezugs auf Modelle der kommunikativen L2-Kompetenz nicht klar, inwiefern der GeRS Aussagen über den Erwerbs*verlauf* treffen kann und will. Unterteilt man die Ausprägung der Kompetenz auf einer Skala in mehrere Stufen, drängt sich die longitudinale Perspektive jedoch geradezu auf. Die Kompetenz wird im GeRS jedoch nicht an Entwicklungsaspekte gekoppelt (Hulstijn 2010: 234). North (2000: 318) konstatiert, dass es keine Garantie

44 Aus Platzgründen kann auf die strategische und die soziolinguistische Kategorie nicht näher eingegangen werden. Die beiden ursprünglich geplanten Kategorien der soziokulturellen (neben der soziolinguistischen) Kompetenz und der Unabhängigkeit (d.h. des Maßes der Selbständigkeit eines Getesteten bzw. der Hilfe, die der Interviewer ihm anbieten muss, um ein Gespräch am Leben zu erhalten) konnten im Laufe des *Schweizer Projekts* nicht beibehalten werden (vgl. Schneider/North 2000: 35).

45 In der vorliegenden Arbeit werden auch deshalb recht umfassend Publikationen zu den Konstrukten (des Lernerwortschatzes und der Flüssigkeit) rezipiert und dokumentiert, weil das mit Bezug auf die GeRS-Skalen bislang nicht erfolgt ist. Dass eine wenn auch nicht vollständige, so doch umfassendere Modellierung einiger Komponenten möglich ist, wird am Beispiel des Arbeitsmodells der lexikalischen Kompetenz versucht zu zeigen.

dafür geben könne, dass die GeRS-Skalen den Erwerbsverlauf spiegelten. Gleichzeitig ist zu berücksichtigen, dass

> "(…) the fact that the CEFR levels are referred to as "natural levels" (…) it is not surprising that the level descriptors in the CEFR are mistakenly understood to reflect discrete stages of language acquisition by practitioners in Europe" (Fulcher 2008: 169).

Auch wenn die GeRS-Autoren also die Deskriptoren lediglich als Hilfsmittel verstehen, um Lernersprache zu beschreiben, werden die Niveaustufen häufig als hierarchische Erwerbssequenz missverstanden.

(6) Zudem ergeben sich *Kohärenzprobleme*, wenn man die Skalen im Kontext des Referenzrahmens betrachtet. Nicht nur ist in den Skalierungs-Publikationen stets von 'proficiency' die Rede, während im Text des Referenzrahmens von 'competence' gesprochen wird:

> "In the Council of Europe Framework, a distinction has been made between the competence categories used in the descriptive scheme in the Framework and the proficiency categories used in the illustrative scales of descriptors for the Common Reference Levels. The latter are derived from the study in this book" (North 2000: 53).

Vielmehr stehen Text und Skalen oftmals unverbunden nebeneinander, wird der Skaleninhalt nahezu nie diskutiert (vgl. Wisniewski 2012a). Teilweise existieren Skalen, die im Text des Dokuments lediglich in einem Nebensatz erwähnt werden (z.B. die Flüssigkeitsskala). Im Text finden sich keine Informationen darüber, was auf welchem Niveau erwartet werden kann; auch Fulcher (2004: 259) moniert die Trennung zwischen GeRS-Text und den Skalen. Fulcher kritisiert ferner, dass der Text des GeRS sich intuitiv auf die so genannten *Threshold*-Dokumente bezöge (Europarat 1975, 1994b [1981], 1999 [1980]), während die Deskriptoren selbst gänzlich ohne Theorie auskämen (2004: 259).

Zusammenfassend lässt sich konstatieren, dass trotz des Bemühens um eine theoretische Ankoppelung an gängige Modelle der Sprachkompetenz die Methode der Skalierung nicht zu Unrecht als untheoretisch bezeichnet wurde (Fulcher 2003: 112; 2004: 258). Eine gewisse Nähe etwa zu Bachman 1990 oder Canale 1983 entsteht durch den Versuch der Übernahme deren Taxonomien, die aber inhaltlich nicht reflektiert und angepasst werden. Bestehende Deskriptoren werden im Nachhinein zu den so etablierten Kategorien sortiert, nicht jedoch aus ihnen abgeleitet und auf sie abgestimmt. Die hier vorgebrachten allgemeinen Kritikpunkte 1–5 werden in Kapitel 2.1.2. und 2.2.2 durch spezifische Betrachtungen zu wortschatz- und flüssigkeitsbezogenen theoretischen Aspekten vertieft.

1.2.3.2 Die empirische Validität der GeRS-Skalen: eine monodimensionale Perspektive

Die empirische Validität der GeRS-Skalen ist als lückenhaft zu betrachten. Sie liegt ganz allein auf den Schultern der Wahrnehmung der Bewerter, die im *Schweizer Projekt* den ihnen vorgelegten Deskriptoren Kategorien (vgl. Kapitel 1.2.3.1) und Schwierigkeitsstufen zuordneten. In intuitiven und qualitativen Verfahren entschied letztlich die Metasprache der Lehrenden über die Kategorisierung der Deskriptoren (vgl. Harsch 2005: 181). Die vertikale Sortierung der Deskriptoren wiederum wird durch das Multifacetten-Rasch-Verfahren als validiert betrachtet (vgl. etwa North 2007b); dieses beseitigt jedoch nicht den Mangel an zugrunde liegender empirischer Validität, sondern bezieht sich erneut hauptsächlich auf die Konsistenz des Verhaltens der Bewerter, ohne einen Bezug zu empirischen Lernerdaten herzustellen.

Der Einbezug von Praktikern ist zweifelsohne von großer Bedeutung – schließlich werden Lernerleistungen sehr häufig gerade von diesen Bewertern eingeschätzt. Allerdings ist im Falle des GeRS dadurch nicht nur die theoretische Basierung, sondern auch, wie im Folgenden dargestellt wird, die Koppelung an die Empirie unter den Tisch gefallen. Es ist nämlich in keiner Weise bekannt, ob das, was die Deskriptoren beschreiben, auch tatsächlich einen Bezug zu empirischer Lernersprache hat, m.a.W. ob sich Lernersprache mit Hilfe der GeRS-Skalen überhaupt erfassen lässt. Das einzig Empirische an der Deskriptorenskalierung ist ihre Bindung an Bewerterurteile:

> "The CEFR is good for Europe and its citizens. However, we must not forget that its empirical base consists of judgments of language teachers and other experts with respect to the *scaling of descriptors*. (…) It is crucial to note, however, that the CEFR is not based on empirical evidence taken from *L2-learner data*". (Hulstijn 2007: 7, Hvhbg. im Orig.).

Zu keinem Zeitpunkt der Skalierung wurde überprüft, ob Lerner tatsächlich (hauptsächlich) sprechen und (seltener) schreiben, wie es die Deskriptoren vorhersagen. Stattdessen wurde einer (Bewertungs-)Konvention - auch durch die Kalibrierung der Deskriptoren mit Hilfe der Multifacetten-Rasch-Analyse - im Nachhinein der Anschein eines objektiven Faktums verliehen. Fulcher/Davidson (2007: 232) sprechen bezüglich der ihrer Ansicht nach bei großen Mess-Unterfangen nicht seltenen Neigung, ein abstraktes Konzept zu einem harten Fakt zu machen, von *Reifizierung*.

Nicht nur von Seiten der Angewandten Linguistik und der Testwissenschaft ist explizit diese Tatsache in den letzten Jahren einige Male kritisiert worden (vgl.

z.B. Tschirner 2005a: 55 und Little 2007: 648; insgesamt Fulcher 2004 und Hulstijn 2007); Harsch (2005: 213) bezeichnet die Deskriptoren gar als 'hypothetisch'. Selbst die Verantwortlichen des *Schweizer Projekts* waren sich dieser Einschränkung ihres Ansatzes prinzipiell bewusst:

> "(…) the difficulty of a descriptor on the scale will be fixed **in relation to a convention** in terms of how it is interpreted. But it means that that convention will based [sic] upon a relatively wide and consistent consensus, rather than copied unthinkingly from an existing scale, and that that conventional interpretation will be objectively calibrated" (North 2000: 38, Hvhbg. K.W.).

Brian North ist vollkommen klar, dass die Skalen des GeRS die objektive Skalierung eines subjektiven Konsens darstellen (ders.: 71). Genauso gesteht er zu, dass die Skalen nicht unbedingt Aussagen ermöglichen über

> "(…) the nature or structure of proficiency, even of the group of learners involved. The most that can be said is that it maps that proficiency with objective scale values in terms of categories, perceptions and conventions which are shared by the group of raters concerned". (ebda.)

Die eingeschränkte empirische Validität der GeRS-Skalen kann nicht ausschließlich der Projekt-Methode der Skalierung angelastet werden, denn die Validitätsprobleme resultieren vor allem aus einem *missverstandenen Gebrauch* der Skalen. Es ist ja durchaus nachvollziehbar, dass, auch aus Praktikabilitätsgründen, im *Schweizer Projekt* nicht alle empirischen Perspektiven (vgl. Drei-Säulen-Konzept, Kapitel 1.3.2.2) gleichermaßen abgedeckt werden konnten. Problematisch ist vor allem die Selbstverständlichkeit, mit denen den GeRS-Skalen heute allgemein eine Aussagekraft zugestanden wird, die über das, was sie eigentlich können und auszusagen beabsichtigten – Bewertungskonventionen von Deutsch-, Englisch- und Französischlehrern in der Schweiz spiegeln – enorm hinausgeht. Nutzer

> "(…) are beginning to believe that the scales in the CEF represent an acquisitional hierarchy, rather than a common perception." (Fulcher 2004: 260)

Ein solches Verständnis der GeRS-Skalen ist aber nicht abwegig, denn Nutzer des GeRS müssten sich schon vertieft mit der Skalierungsmethode auseinandersetzen, um sich darüber bewusst zu werden, wo die Einschränkungen der Skalen liegen. Durch die Platzierung und Kommentierung der Skalen im Referenzrahmen und die selbstverständliche Forderung nach ihrer Anwendung im Lehr-, Lern- und Testalltag wird eine überhöhte Erwartung an die Instrumente befördert.

Ob man mit den Skalen wirklich beschreiben kann, was Lerner (in Tests) für Sprache produzieren, müsste in einem groß angelegten korpusbasierten empirischen Projekt erst noch untersucht werden. Dabei ist Hulstijn (2007: 9) Recht in

seiner Ansicht zu geben, dass ein Instrument wie der Referenzrahmen es sich nicht leisten kann, keine empirische Überprüfung der Skalen durchzuführen:

> "Educational authorities and politicians must be made aware of the missing linguistic, psycholinguistic and sociolinguistic poles underneath the CEFR and the urgency of making funds available for collaborative research. To provide a building with a proper foundation, it need not be tore down first and built up later. The educational community can continue to reside in the CEFR, while researchers are constructing the poles underneath."

Bislang gibt es keine Nachweise dafür, dass die Deskriptoren Lernersprache überhaupt beschreiben können, dass man wirklich schrittweise durch die Niveaustufen geht und auch andere Aspekte bleiben unbewiesen (Hulstijn 2007: 8). Es gibt keinerlei korpusbasierte Studien, die die Validität der GeRS-Skalen in dieser Hinsicht untersuchen würden.[46] Wenn Deskriptoren aussagekräftig sein wollen, müssen sie aber diese Verbindung zur Realität der Lernersprache herstellen können:

> "If descriptors are to be meaningful characterizations of ability, then they should be able to be related to actual performance." (Alderson 1991: 74)[47]

1.3 Das Design dieser Arbeit

1.3.1 Forschungsfragen

Die Forschungsfragen ergeben sich aus dem gerade dargelegten Validitätsdefizit der GeRS-Skalen und beziehen sich auf die drei Säulen der Skalenvalidität, die unten weiter ausgeführt werden (vgl. Kap. 1.3.2):

(1) Sind die Skalen für Flüssigkeit, Wortschatzspektrum und Wortschatzangemessenheit des GeRS theoretisch kohärent?

---

46 Auch wenn Hulstijn, Alderson und Schoonen (2010: 16) wie auch die Autorin der vorliegenden Studie die Skalen des fünften Kapitels für am problematischsten halten, sind doch Ansätze wie der von den drei Linguisten vertretene nicht dazu geeignet, die Skalen des GeRS selbst auf ihre Validität hin zu validieren. In den Studien der SLATE-Gruppe nämlich (siehe oben) werden GeRS-Bewertungen mit Charakteristika von Lernersprache abgeglichen. Dies ist zwar hochinteressant, sobald man sich zur Akzeptanz der Existenz der GeRS-Niveaus durchgerungen hat. Andererseits entsteht hier eine methodische Zirkularität, denn wieder können so Bewertereindrücke, auf deren Basis dann zugehörige linguistische Charakteristika erst zugeordnet werden, wissenschaftliche Studien entscheidend mitbestimmen.

47 Vgl. ähnlich Figueras/North/Takala/Verhelst/VanAvermaet 2005: 271 und Fulcher 2008: 173.

(2) Kann man mit den Skalen für Flüssigkeit, Wortschatzspektrum und Wortschatzangemessenheit des GeRS italienische und deutsche Lernersprache, wie sie in einem mündlichen Sprachtest produziert wurde, auf den Niveaustufen A2, B1 und B2 angemessen beschreiben? (siehe Unterfragen 2.1–2.3)
(3) Ist es plausibel, dass Bewertungen dieser Lernerleistungen tatsächlich auf den operationalisierten Deskriptoren beruhen?

Die erste Forschungsfrage tangiert das Problem der theoretischen Kohärenz der gewählten GeRS-Skalen; auch wenn die Skalen, wie oben dargelegt wurde, nicht aus einer Theorie bzw. einem Modell abgeleitet wurden, ist durchaus denkbar, dass die skalierten Deskriptoren ein kohärentes Bild ergeben, das sich einer oder mehreren Theorien zuordnen lassen könnte. Es wäre dann *ex post* ein stimmiger Theoriebezug entstanden. Außerdem kann hier geprüft werden, ob und wie zentral die Deskriptoren tatsächlich zum Konstrukt der Flüssigkeit, der Wortschatzangemessenheit und der Wortschatzbreite gehören bzw. welche Dimensionen dieser Konstrukte sie abzudecken vermögen.

Die zweite und dritte Forschungsfrage sind empirischer Natur; die zweite steht im Zentrum der Arbeit. Sie stellt die oben aufgeworfene Frage nach der Passung von Skaleninhalten und Lernersprache und konkretisiert sich in drei Unterfragen:

(2.1) Kann man die zu *Skalenvariablen* operationalisierten Deskriptoren der genannten Skalen stabil, d.h. unabhängig von der Zielsprache (Italienisch/Deutsch) und dem Aufgabentyp (monologisch/dialogisch), beobachten? (*Kriterium der empirischen Relevanz*)
(2.2) Kann man Lernerleistungen sinnvoll unabhängig von Aufgabentyp und Sprache gruppieren, wenn man sich allein auf die operationalisierten Deskriptoren verlässt, und lassen sich die so gruppierten Leistungen deutlich von anderen Leistungen unterscheiden? (*Kriterium der Konvergenz/Trennbarkeit*)
(2.3) Inwiefern ist es möglich, die operationalisierten Deskriptoren empirisch in Zusammenhang mit etablierten Maßen des jeweiligen Konstrukts (Flüssigkeit bzw. Wortschatzkorrektheit und –breite) zu bringen? (*Kriterium des empirischen Konstruktbezugs*)

Frage (2.1) bildet die Basis für alle folgenden Forschungsfragen: Wenn das, was in den Deskriptoren beschrieben wird, empirisch nicht, nicht häufig genug oder nicht unterschiedlich genug ausgeprägt vorkommt, entstehen Probleme für die Verwendung der Niveaubeschreibung zur Einschätzung empirischer Lernersprache, denn der entsprechende Deskriptor bietet dann keine Hilfestellung für Entscheidungen.

In Frage (2.2) geht es darum, ob die in den Niveaubeschreibungen jeweils erwähnten Phänomene in der dort *vorhergesagten Ausprägung* (also zum Beispiel: *viele* Pausen, *wenig* Fehlstarts) bei Lernern *gemeinsam* auftreten. Nur dann lassen

sich Lernerproduktionen mit den Deskriptoren beschreiben. Wenn das der Fall ist, wäre wünschenswert, dass man Lernerleistungen, die zu einer operationalisierten Niveaubeschreibung passen, deutlich von anderen Leistungen trennen kann, bei denen das nicht der Fall ist.

Frage (2.3) zielt darauf ab, Nachweise dafür zu finden, dass die operationalisierten Deskriptoren (so genannte *Skalenvariablen*) tatsächlich die jeweiligen Konstrukte messen, indem sie an übliche Maße der Flüssigkeit, der Wortschatzangemessenheit und der Wortschatzbreite gekoppelt werden. Sollte kein Zusammenhang der Niveaubeschreibungen mit Maßen der Konstrukte gefunden werden, liegt der Verdacht nahe, dass die Skalen Konstruktirrelevantes messen. Frage (2.3) liegt an der Schnittstelle von Theorie und Empirie.

Frage (3) stellt Aspekte der Bewertbarkeit, denen ja bereits bei der Skalierung der GeRS-Skalen große Aufmerksamkeit zukam, in den Mittelpunkt, allerdings aus einer etwas anderen Perspektive. Hier wird danach gefragt, ob anzunehmen ist, dass die am Dissertationsprojekt beteiligten Bewerterinnen für ihre Urteilsbildung ausschließlich oder zumindest wesentlich auf die Inhalte der Niveaubeschreibungen (A2, B1, B2) der drei Skalen zurückgegriffen haben. Sollte sich herausstellen, dass Skalenvariablen eine geringe Rolle für die Urteilsbildung gespielt haben, stellt sich die Frage nach etwaigen skalenexternen, möglicherweise gar konstruktfremden Einflussfaktoren.[48]

---

48 Der Aspekt der theoretischen Fundiertheit stellt die Anforderung an die Skala, an ein Modell der Kompetenz gekoppelt zu sein und Ergebnissen der L2-Erwerbsforschung nicht zu widersprechen; dies ist im Grunde eine Frage der Konstruktvalidität, von Bedeutung auch bei Bachman/Palmer (1996). Das Kriterium der Authentizität lässt sich nicht ohne Weiteres übernehmen (Bachman/Palmer 1996: 23). Bei *real-life*-Skalen wäre es relativ nahe liegend, nach dem Realitätsbezug der Aufgaben zu fragen, auf die sich die Skalen beziehen; die Inhalte der Deskriptoren müssen zu den Aufgaben passen (Alderson 1991: 74). Bei *ability*-Skalen hingegen ist das nicht sinnvoll, weil sie aufgabenunabhängig funktionieren müssen. Der kritische Punkt liegt hier vielmehr darin, dass trotz der Kontextfreiheit der Skalen jede Aufgabe, die zur Beurteilung steht, mit der Skala auch beurteilt werden können muss. Gerade bei den Wortschatzskalen wird sich zeigen, dass nicht mit jeder Art von Aufgabe ein so genanntes 'ratable sample' erhoben werden kann (Bachman/Palmer 2010: 351, vgl. Kapitel 3.2.4). Diese – versteckte – Aufgabenabhängigkeit hat enorme Konsequenzen auf die komplette Testprozedur. Bewertbare Lernerproduktionen müssen den Kandidaten die Gelegenheit geben, die volle Bandbreite an Aspekten zu demonstrieren, die bewertet werden sollen (ebda.); dies ist eine Herausforderung nicht nur in horizontaler, sondern auch in vertikaler Hinsicht. Die Frage danach, ob mit einer Skala *ratable samples* erhoben werden können, hängt auch an den Formulierungen in der Skala selbst: sind diese sehr

1.3.2 Validierungsansatz

Im Folgenden soll der hier verwendete Validierungsansatz dargelegt werden. Dabei wird zunächst die Möglichkeit einer Validierung von kontextfreien *Ability*-Skalen thematisiert, wobei verdeutlicht werden soll, dass das Ziel der vorliegenden Studie keine vollständige Validierung isolierter Skalen sein soll, sondern eine *Abklärung der Möglichkeiten des validen Einsatzes dieser Instrumente*. Im Anschluss wird der auf drei Säulen basierende Validierungsansatz vorgestellt.

Da Skalen i.d.R. nicht isoliert auf ihre Validität untersucht werden, erzwingt das Vorgehen eine modifizierte Verwendung der in der Testwissenschaft **üblichen Terminologie**. Dennoch lassen sich die drei genannten Eckpfeiler teilweise auf Bachman/Palmers (1996, 2010) Gütekriterien von Sprachtests zurückführen.

1.3.2.1 Denkrahmen: Die Abklärung der Möglichkeiten des validen Skaleneinsatzes

Eine Untersuchung der Validität von Skalen, die nicht nur Aussagekraft hinsichtlich eines konkreten, sondern mehrerer möglicher Testkontexte erlangen möchten, muss konzipiert werden als eine *Analyse der Möglichkeiten, mit Hilfe der Skalen zu Bewertungen zu kommen, die die Grundlage für angemessene Handlungen und Interpretationen bilden können*. Damit werden die Skalen hier aus dem ‚üblichen', kontextgebundenen Validierungsprozess extrahiert und auf einige als zentral erachtete Aspekte hin untersucht (vgl. Kapitel 1.3.2). Die Tatsache, dass die GeRS-Skalen für konkrete Testkontexte erst angepasst werden

---

spezifisch, kann es möglich sein, dass für jede Niveaubeschreibung oder im Extremfall sogar für jeden Deskriptor die Konstruktion einer eigenen Testaufgabe nötig wäre (vgl. Kapitel 3.2.4). Dann ist aber höchst fraglich, inwiefern solche Deskriptoren Aussagen über eine zugrunde liegende Kompetenz ermöglichen sollen. Als eine gewisse Art von Authentizität lässt sich jedoch der Lernerbezug bewerten, der in dieser Studie prominente Stellung genießt. Ebenso ist es eine Frage der Fairness, die Leistungen von Lernern nur mit Instrumenten beurteilen zu wollen, die Aussagen darüber treffen können, was Lerner tatsächlich tun. Während die Gütekriterien der Reliabilität, der Interaktivität und des Impact in dieser Arbeit nicht untersucht werden, spielt die Praktikabilität eine Rolle: hierunter wird in diesem Kontext die Frage danach verstanden, inwiefern Bewerter ihre Urteile durch den Rekurs auf die Skalen fällen können bzw. auf andere Entscheidungshilfen angewiesen sind (vgl. Kapitel 1.3.2.2). Etwas außerhalb der Bachman/Palmer Schen Gütekriterien steht die hier erhobene Forderung nach Kohärenz: Skalen können nur dann nützlich sein, wenn die Niveaubeschreibungen empirisch und theoretisch widerspruchsfrei sind.

müssen, enthebt sie nicht der Pflicht, wenn auch lückenhafte, Nachweise für die *Möglichkeit* ihres validen Einsatzes zu bringen. Sind diese Nachweise erbracht, kann eine Skala in verschiedenen Testkontexten eingesetzt werden, die dann valide Ergebnisinterpretationen ermöglichen (aber nicht erzwingen!). Sind die Kriterien hingegen nicht zufrieden stellend erfüllt, muss davon ausgegangen, dass die Validität der Ergebnisinterpretation in konkreten Testsituationen durch die Verwendung der entsprechenden Skala bereits *strukturell* bedroht wird. Der Hinweis darauf, dass die Skalen in einem konkreten Testkontext erst modifiziert werden müssten, und ohnehin für jeden Einsatz als Bewertungsinstrument verändert werden müssten, tut der Tatsache keinen Abbruch, das das, was sich in den Skalen in ihrer jetzigen Form findet, bereits sowohl einen Empirie- als auch einen Theoriebezug braucht. Eine wesentliche, strukturverändernde Modifikation der Skaleninhalte würde auch die Frage aufwerfen, inwiefern eine angepasste Skalenversion auf ihre Vorlage überhaupt noch angewiesen ist.

1.3.2.2 Der Drei-Säulen-Ansatz

Als zentrale Charakteristika für die Schaffung der Möglichkeit eines validen Skaleneinsatzes werden (1) die Bindung der Skala an Theorien oder Modelle fremdsprachlicher Kompetenz, (2) die Beziehbarkeit der Skaleninhalte auf authentische Lernerproduktionen und (3) auch die Handhabbarkeit der Skalen durch Beurteilende verstanden. Der letztgenannte Punkt bezieht sich weniger auf die Reliabilität, die nichts darüber aussagt, wie stark Beurteilende bei ihrer Arbeit tatsächlich auf die Skala rekurrieren, sondern vor allem auf Analysen dazu, inwiefern sich Bewerterurteile tatsächlich durch die verwendeten Skalen erklären lassen.

Mit diesen drei Säulen (vgl. Abb. 1), hier untersucht in den oben dargestellten Forschungsfragen 1–3, wird selbstverständlich nicht jeder erdenkliche Validitätsaspekt behandelt, sondern ein Schwerpunkt auf als besonders zentral begriffene Themen gelegt.

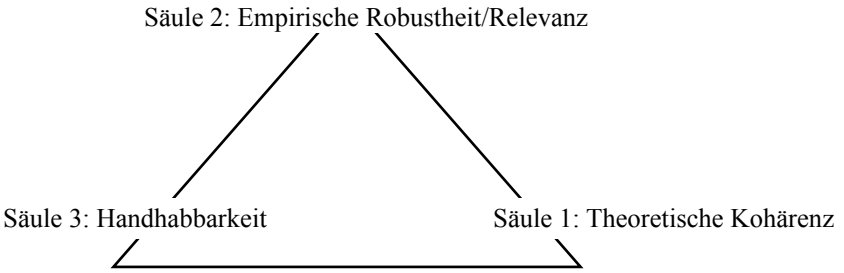

Abb. 1: Voraussetzungen für einen validen Einsatz der GeRS-Skalen

Es wird also keine vollständige Validierung (auch nicht der drei ausgewählten Skalen) angestrebt. Dafür wäre nicht nur der Einsatz weiterer Auswertungs-Methoden, sondern auch eine erheblich größere Stichprobe vonnöten gewesen, ebenso wie der Einbezug weiterer Sprachen.

Säule 1: Theoriebezug

Der Ansatz richtet sich gegen die Akzeptanz von Skalen als Quasi-Konstrukte. *Ability*-Skalen (und als solche müssen die Skalen aus dem fünften Kapitel des Referenzrahmens verstanden werden, vgl. Kapitel 1.2.2.1) bedürfen des gültigen Bezugs auf ein zugrunde liegendes Modell der kommunikativen L2-Kompetenz (auf die ja abstrahiert werden soll),[49] das wiederum in Einklang mit den Erkenntnissen der Zweit- bzw. Fremdsprachenerwerbsforschung zu stehen hat. Dafür ist es nicht ausreichend, die Kategorien zur Skalenbildung an die Taxonomien verschiedener gängiger Modelle anzunähern, wie im *Schweizer Projekt* geschehen (vgl. Kapitel 1.2.3.1), sondern es muss ein Nachweis dafür erbracht werden, dass die Skalen *inhalte* theoretisch relevant und kohärent sind. Das Aufgreifen einer bloßen Taxonomie lässt zudem den kompliziertesten, aber auch wichtigsten Teil von Modellen außen vor, nämlich den funktionalen Zusammenhang der Modellkomponenten.

Natürlich kann eine Skala, die Aussagen zur Kompetenz ermöglichen will, nur sehr allgemeine Beschreibungen enthalten, die losgekoppelt von konkreten Aufgaben (kontextfrei) sein müssen. Nichtsdestotrotz müssen die enthaltenen Deskriptoren dennoch kohärent und bedeutsam für das zu messende Konstrukt sein.

Auch der Verweis auf bislang unvollständige Erwerbstheorien (vgl. z.B. Kapitel 1.2.3.1) ist kein Grund, einen solchen Theorie-Validierungsschritt zu unterlassen bzw. zu erstellenden Skalen keine theoretischen Analysen vorauszuschicken. Trotzdem liegen hier einige gravierende Schwierigkeiten, denn die vorliegenden Modelle der kommunikativen Kompetenz sind nicht ausführlich genug, um auf ihrer Grundlage aussagekräftige Bewertungsskalen zu entwickeln (vgl. Knoch 2011: 85). In dieser Arbeit hat sich beispielsweise gezeigt, dass die Flüssigkeit in solchen Modellen gar nicht, der Wortschatz nur in sehr undifferenzierter Weise erwähnt wird. Obwohl die Flüssigkeit ein sehr beliebtes Bewertungskriterium ist, tritt sie also in den Modellen gar nicht auf. Für die lexikalische Kompetenz wurde hier ein Arbeitsmodell entwickelt (vgl. Kapitel 2.2.2.6), ein aufwändiger, sicherlich nicht in jedem praktischen Zusammenhang gangbarer Weg. Ein weiteres

49 Oder zur *ability* oder *proficiency*; das Prinzip ist stets gleich, wenn von einer konkreten Performanz auf eine allgemeinere zugrunde liegende Fähigkeit geschlossen werden soll.