Die Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen im Praxistest: Eine empirische Studie zur Validität des Referenzrahmens

Olaf Bärenfänger[1]

Since its publication in 2001, the Common European Framework of Reference for Languages (CEFR) has had a tremendous impact on the teaching and learning of foreign languages. Particularly in the field of language testing, the major European language tests rely on the CEFR system of competence scales. Although the CEFR has been criticized on theoretical grounds from the beginning, relatively little research has addressed the question to which extent CEFR scales are suitable to evaluate authentic learner productions. This paper presents an overview of the research literature relating to the validity of the CEFR scales. The article focuses on a huge sample of written German and Italian speech productions that was analyzed statistically with respect to scale functionality. The underlying texts had been rated based on seven CEFR scales. Results indicate that the CEFR scales under inspection are not fully functional. Therefore, more research is suggested that relates the theoretical framework of the CEFR to authentic learner language. By that means, a revision of the CEFR might mitigate some of the flaws identified in this study.[2]

1. Einleitung[3]

Der Gemeinsame europäische Referenzrahmen für Sprachen (GeR) wurde mit dem Anspruch publiziert, durch die Formulierung von sprachlichen Kompetenzstandards zu einer besseren Vergleichbarkeit der Bildungssysteme in Europa beizutragen (vgl. den Beitrag von Neil Jones in diesem Heft). Mit der Schaffung eines weithin akzeptierten Maßstabs für unterschiedliche Ausprägungen fremdsprachlicher Kompetenzen sollten Planungen in allen Bereichen des Fremdsprachenunterrichts optimiert werden. Zugleich sind einheitliche Leistungsmaßstäbe eine notwendige Voraussetzung für die Vergleichbarkeit von Sprachprüfungen. Schließlich verfolgt der GeR auch das Ziel, selbstbestimmtes Lernen zu fördern (vgl. zu all diesen Punkten Europarat 2001: 16).

1 Priv.-Doz. Dr. Olaf Bärenfänger, Goethestr. 2, 04109 Leipzig, Tel. + 49 341 97-30270, baerenfaenger@uni-leipzig.de

2 Die im Rahmen der vorliegenden Studie analysierten Daten waren mit einer anderen Zielsetzung bereits Gegenstand in folgendem technischen Bericht: Olaf Bärenfänger (2013): *Assessing the Reliability and Scale Functionality of the MERLIN Written Speech Sample Ratings*. Technical Report. Leipzig: University of Leipzig.

3 Ich danke den Herausgebern, zwei anonymen Gutachtern sowie Herrn Jupp Möhring und Frau Katrin Wisniewski für zahlreiche wertvolle Hinweise und Frau Hedwig Rebekka Koch für ihre Unterstützung bei der Formatierung des Beitrags.

Angesichts der eminenten Auswirkungen, die der GeR auf das gesamte Bildungssystem, aber auch auf individuelle Bildungsverläufe hat, drängt sich die grundsätzliche Frage auf, inwiefern sich die Skalen des GeR überhaupt bei der aussagekräftigen Beschreibung fremdsprachlicher Kompetenzen bewähren. Terminologisch formuliert stellt sich somit das Problem der Validität von GeR-Skalen. Von seiner Entstehung her bildet der GeR lediglich die statistisch aggregierten Einschätzungen einer großen Zahl von Lehrenden ab, welche fremdsprachlichen Kompetenzen Lernende auf einem bestimmten Niveau aufweisen (sollen). Inwiefern diesen naturgemäß subjektiven Meinungen objektive, empirisch beobachtbare Sachverhalte entsprechen, wurde bislang nur in Ansätzen erforscht (vgl. dazu Wisniewski 2014).

Vor diesem Hintergrund widmet sich der vorliegende Beitrag der Frage, inwiefern sich die Kompetenzskalen des GeR bei der Bewertung schriftlicher Lernendenleistungen in der Praxis als funktional erweisen. Als funktional wird eine GeR-Skala dann verstanden, wenn sie bei der Anwendung auf authentische Lernersprache plausible, erwartungstreue Ergebnisse hervorbringt; eine Lernerproduktion auf dem Niveau B1 müsste also z. B. überwiegend mit B1-Deskriptoren beschrieben werden. Ferner sollten funktionale Skalen für fremdsprachliche Kompetenzen einen inhaltlichen Zusammenhang aufweisen; dieser sollte u. a. in Korrelationen zwischen den Einzelskalen erkennbar werden. Schließlich sollten funktionale Skalen gleiche Schwierigkeitsgrade widerspiegeln; die Schwierigkeit von Skalen lässt sich mithilfe von Mulitfacetten-Rasch-Analysen ermitteln.

Im folgenden Abschnitt erfolgt zunächst eine kritische Darstellung des GeR mit Fokus auf der Validität seiner Skalen. Im Anschluss daran wird im dritten Abschnitt die Funktionalität von sieben exemplarisch ausgewählten GeR-Skalen untersucht, die auf ein großes Korpus authentischer Lernersprache angewendet wurden. Die empirische Studie basiert auf den Daten von 898 Lernenden des Italienischen, die einen offiziellen standardisierten *high-stakes test* auf den Niveaus A1, A2, B1 oder B2 abgelegt hatten, sowie auf den Daten von 1.139 Lernenden des Deutschen auf den Niveaus A1, A2, B1, B2 oder C1. Die schriftlichen Produktionen dieser Testteilnehmenden waren von je zwei professionellen, geschulten Bewertenden nachbewertet worden. Zur Beurteilung der Bewerterqualität wurden ca. 10% der Aufsätze doppelt bewertet. Häufigkeitsanalysen der vergebenen Niveaustufen, Korrelationen zwischen den Skalen sowie Multifacetten-Rasch-Analysen erlauben detaillierte Rückschlüsse auf die Qualität der Skalen. Ein Abschnitt mit Schlussfolgerungen und mit der Benennung einiger Forschungsdesiderate beschließt den vorliegenden Beitrag.

2. Die Validität der GeR-Skalen im Spiegel der Forschung

Bereits von Anbeginn an sah sich der GeR massiver Kritik ausgesetzt, die sich auf das Skalensystem und die Anwendbarkeit der Skalen auf Lernersprache richtete. So wurde dem GeR neben einer geringen Benutzerfreundlichkeit (Hilpisch 2012), eine teils fehlerhafte Übersetzung (Harsch 2005) sowie eine unklare Terminologie unterstellt (Cools & Sercu 2006, Alderson 2007, Figueras 2012). Mit ähnlicher Stoßrichtung wurde die mangelnde Präzision der Leistungsdeskriptoren bemängelt (vgl. z. B. Barkowski 2003, Quetz 2003, Cools & Sercu 2006, Hilpisch 2012, Kusseling & Lonsdale 2013 oder Simons & Colpaert 2015). Publikationen wie Weir (2005) oder Cambridge ESOL (2011) bescheinigen dem GeR sogar Miss- bzw. Unverständlichkeit. Der notwendigerweise allgemeine Charakter der GeR-Skalen stellt für Harsch (2005) ein Hindernis bei ihrer Verwendung in konkreten Kontexten dar. Harsch & Martin (2012) hatten deshalb bereits auf die Notwendigkeit hingewiesen, GeR-Skalen anzupassen, wenn diese auf Lernerproduktionen in Tests angewendet werden sollen.

Wisniewski (2014) kommt zu dem Schluss, dass für keine der von ihr untersuchten Skalen "Flüssigkeit", "Wortschatzspektrum" und "Angemessenheit der Verwendung des Wortschatzes" durchgängig klar sei, inwiefern die Skaleninhalte an die aktuelle Forschung anschlussfähig sind (siehe zur ungenügenden theoretischen Fundierung des GeR bereits Vogt 2011 und House 2003). Ein Anschluss an die Ergebnisse der Fremdsprachenerwerbsforschung und Modelle der kommunikativen Kompetenz ist damit zumindest nicht lückenlos, wenn überhaupt, gegeben. Auch schätzt Wisniewski (2014) die Kohärenz der Niveaustufenbeschreibungen als nicht überzeugend ein. Insbesondere bestehe eine Kluft zwischen dem Text des GeR und den Skalen.

Unter Validitätsgesichtspunkten ebenfalls problematisch ist der subjektive Charakter der GeR-Skalen, die lediglich Einschätzungen von Fremdsprachenlehrenden widerspiegeln, welche Kompetenzen ein Lernender auf einem bestimmten Niveau hat (siehe dazu auch Roche 2005, Schneider 2005, Hulstijn 2007 sowie Kusseling & Lonsdale 2013). Diese "gefühlten Skalen" (Quetz 2007: 49) halten einer linguistischen oder lerntheoretischen Analyse oftmals nicht stand. Mit Blick auf die Validität der GeR-Skalen erhebt sich indessen die Forderung, empirische Lernerdaten konsistent modellieren zu können (vgl. in diesem Sinne Hulstijn 2007).

Auf ein weiteres empirisches Defizit weist eine Studie der Europäischen Kommission (2013, vgl. den Beitrag von Neil Jones in diesem Heft) hin, in der die Autoren die Umsetzung des GeR in fünf ausgewählten europäischen Ländern untersucht hatten. Grundsätzlich wurde in diesem Zusammenhang das Fehlen von empirischen Belegen für die Verbindung des GeR mit Lernergebnissen, Zielsetzungen, Prüfungen und politischen Strategien in Europa vermisst.

Bilanziert man die genannte Literatur, lässt sich festhalten: Trotz der zweifellos beispiellosen Wirkungsmächtigkeit des GeR stehen Zweifel an der umfassenden Validität seiner Skalen im Raum. Die nachstehende Studie soll einen Beitrag dazu leisten, die Einwände entweder empirisch zu bekräftigen oder zu widerlegen.

3. Erkenntnisinteresse und Forschungsdesign

Im Rahmen der vorliegenden Studie wurden schriftliche Produktionen, die im Rahmen von verschiedenen *high-stakes tests* der telc gGmbH entstanden waren, in den Sprachen Italienisch und Deutsch von eigens geschulten Bewertenden gemäß den GeR-Skalen "Allgemeines sprachliches Spektrum", "Wortschatzspektrum", "Wortschatzbeherrschung", "Grammatische Korrektheit", "Kohärenz und Kohäsion", "Soziolinguistische Angemessenheit" sowie "Orthographische Korrektheit" bewertet. Die ursprünglichen Bewertungen standen weder den Bewertenden noch dem Autor zur Verfügung. Außerdem ist zu beachten, dass es sich bei den italienischen Skalen um nicht-kalibrierte Übersetzungen handelt. Etwa 10% der Texte wurden doppelt beurteilt, um die Inter-Rater-Reliabilität überprüfen zu können. Die Beurteilungen wurden nach der Datencodierung einer statistischen Analyse unterzogen, die auf die Häufigkeitsverteilung der Referenzrahmenniveaus sowie auf Korrelationen zwischen den Bewertungskriterien fokussierte. Weiterhin wurden für die Bewertungen in jeder Sprache eine Multifacetten-Rasch-Analyse mit den Facetten "Beurteilerstrenge", "sprachliche Kompetenz des jeweiligen Testteilnehmenden" und "GeR-Skala" durchgeführt.

3.1 Der Forschungskontext

Die vorliegende Studie entstand im Kontext des MERLIN-Projekts, in dessen Rahmen ein nach GeR-Stufen kalibriertes Lernersprachenkorpus mit schriftlichen Produktionen in den Sprachen Deutsch, Italienisch und Tschechisch aufgebaut wurde (vgl. Boyd et al. 2014). Tabelle 1 gibt den Teilbereich des MERLIN-Korpus wieder, der für die vorliegende Untersuchung herangezogen wurde.

Wie Tabelle 1 zeigt, enthält das Untersuchungskorpus vergleichsweise große Stichproben. Für jedes Niveau liegt eine ähnliche Zahl an Bewertungen vor. Mit einer Bandbreite von A1 bis B2 für Italienisch bzw. von A1 bis C1 für Deutsch wird zudem ein relativ breites Kompetenzspektrum abgedeckt.

Alle Lernendenproduktionen waren im Rahmen verschiedener Tests der telc gGmbH elizitiert worden (*The European Language Certificates*; vgl. www.telc.net;

detaillierte Informationen zu den Tests siehe auf dieser Webseite). Dementsprechend ist davon auszugehen, dass die Prüfungen unter standardisierten Bedingungen durchgeführt worden waren.

Tabelle 1: Das Analysekorpus

	GeR-Stufe	Sprachtest	Anzahl von Beurteilungen pro Stufe und Sprache	Gesamtzahl der Beurteilungen
Deutsch	A1	telc: Start Deutsch 1	229	1.139
	A2	telc: Start Deutsch 2	228	
	B1	telc: Zertifikat Deutsch	231	
	B2	telc: Deutsch B2	225	
	C1	telc: Deutsch C1	226	
Italienisch	A1	telc: Italiano A1	229	898
	A2	telc: Italiano A2	224	
	B1	telc: Italiano B1	223	
	B2	telc: Italiano B2	222	
			Gesamt	**2.037**

3.2 Bewertungsverfahren

Alle schriftlichen Produktionen in den Sprachen Deutsch und Italienisch wurden für jede der beiden Sprachen von zwei Bewertenden ohne Kenntnis der ursprünglichen Einstufungen nachbewertet. Die Bewertenden waren Muttersprachler und hatten mehrjährige Erfahrung in der Bewertung von Sprachprüfungen. Auch waren sie eigens für die Bewertung des Prüfungskorpus in einem zweitägigen Workshop geschult worden.

Mit Blick auf die Qualität der Bewertungen und damit zugleich auf die Reliabilität der Datengrundlage dieser Studie wurden ca. 10 % der Lernendentexte doppelt beurteilt. Auf diese Weise ist es möglich, gängige Maße der Interrater-Reliabilität für die Bewerterpaare zu berechnen (vgl. dazu Wirtz & Caspar 2002, Eckes 2010). Tabelle 2 gibt mit Kendalls *tau* ein Maß für die Konsistenz der Urteile und mit Cohens *kappa* ein Maß für die zufallskorrigierte Übereinstimmung der Bewertenden wieder.

Tabelle 2: Maße der Interrater-Reliabilität für die Bewertungen im deutschen Subkorpus

	n	Kendalls *tau*	Cohens *kappa*
Allgemeines sprachliches Spektrum	101	,894	,524
Wortschatzspektrum	101	,880	,600
Wortschatzbeherrschung	101	,832	,524
Grammatische Korrektheit	95	,808	,519
Kohärenz und Kohäsion	98	,851	,593
Sozioling. Angemessenheit	98	,780	,512
Orthographische Korrektheit	81	,781	,484

Bemerkung: Alle statistischen Maße sind signifikant auf dem Niveau $p < 0{,}01$. Die Spalte *n* entspricht der Anzahl der Doppelbewertungen.

Tabelle 2 belegt klar und auf breiter Datenbasis, dass die Bewertenden im deutschen Subkorpus ein durchweg hohes Maß an Konsistenz in ihren Bewertungen erreichen. Die Übereinstimmung zwischen den Bewertenden fällt demgegenüber etwas geringer aus. Gleichwohl ist immer noch von einer hohen Qualität der Bewertungen auszugehen.

Tabelle 3: Maße der Interrater-Reliabilität für die Bewertungen im italienischen Subkorpus

	n	Kendalls *tau*	Cohens *kappa*
Allgemeines sprachliches Spektrum	81	,842	,288
Wortschatzspektrum	81	,814	,698
Wortschatzbeherrschung	81	,621	,281
Grammatische Korrektheit	81	,692	,461
Kohärenz und Kohäsion	81	,761	,574
Sozioling. Angemessenheit	82	,591	,363
Orthographische Korrektheit	81	,805	,314

Bemerkung: Alle statistischen Maße sind signifikant auf dem Niveau $p < 0{,}01$. Die Spalte *n* entspricht der Anzahl der Doppelbewertungen.

Tabelle 3 zeigt für das italienische Subkorpus ein in etwa vergleichbar hohes Maß an Beurteilerkonsistenz wie für das deutsche Subkorpus. Die Beurteilerübereinstimmung ist allerdings geringer ausgeprägt als im deutschen Subkorpus. Die hohe Beurteilerkonsistenz zusammen mit der etwas geringeren Beurteilerübereinstimmung deutet auf ein intraindividuell stabiles Bewerterverhalten bei gleichzeitigen Strengeunterschieden hin. Die Qualität der Bewertungen darf angesichts

der hohen Bewerterkonsistenz immer noch als gut gelten, zumal wenn Strengeunterschiede wie in der vorliegenden Studie statistisch herausgerechnet werden können. Insgesamt dürfen sowohl die Bewertungen der deutschen als auch der italienischen Lernerproduktionen für sich ein hohes Maß an Reliabilität beanspruchen und damit als verlässliche Datengrundlage für die vorliegende Studie gelten.

3.3 Statistische Auswertungen

Bezüglich der Datenreliabilität, also der Qualität der Bewertungen, wurden als gängiges Konsistenzmaß Kendalls *tau* und als zufallskorrigiertes Übereinstimmungsmaß Cohens *kappa* berechnet. Mit Blick auf die Funktionalität der Skalen wurden zur Einschätzung, inwieweit die untersuchten Skalen plausibel und erwartungstreu das intendierte Niveau des Tests abbilden, pro Sprache und Niveaustufe für jedes GeR-Kriterium die relative Häufigkeit der Niveaustufenzuweisungen ermittelt. Bezüglich eines kohärenten Zusammenhangs zwischen den Kriterien wurden pro Sprache und Niveaustufe Spearman-*rho*-Korrelationen zwischen den GeR-Skalen berechnet. Mit Blick auf die Schwierigkeit der untersuchten Skalen wurden Multifacetten-Rasch-Analysen durchgeführt. Neben objektiven Informationen über die Kompetenzausprägung bei den Testteilnehmenden ermöglichen Rasch-Analysen nämlich auch differenzierte Aussagen z. B. über Aufgabenschwierigkeit oder Skaleneigenschaften, was diese statistische Technik besonders interessant für die Validierung von Tests oder Skalen macht. Je nach dem, welche Datentypen betrachtet werden sollen, stehen unterschiedliche Raschmodelle zur Verfügung (Linacre 2009, Eckes 2015). In der vorliegenden Studie wurde drei Facetten der Testsituation berücksichtigt: der Grad der Kompetenzausprägung bei den Testteilnehmenden, die Beurteilerstrenge sowie die Beurteilungskriterien. Im Rahmen dieses Modells mit drei Facetten kann die Ausprägung der fremdsprachlichen Kompetenz bei Testteilnehmenden somit in Abhängigkeit der Beurteilerstrenge und der Bewertungskriterien bestimmt werden. Zusätzlich wurde für jede Facette die sog. Modellpassung berechnet, die als Indikator für die Validität der jeweiligen Daten interpretiert kann. Während ein Wert von 1,0 einem idealen Verhältnis von vorhergesagter und beobachteter Varianz entspricht, deuten im Falle von Rating-Skalen Werte von größer als 1,4 auf eine unbrauchbare Messung hin (Bond & Fox 2007). Die sog. Separationsreliabilität gibt für jede Facette an, inwieweit die Elemente dieser Facette sich voneinander unterscheiden.

4. Ergebnisse

Tabelle 4 gibt für die deutschen Lernendenproduktionen die relative Häufigkeit der Niveaustufenzuweisungen für jedes getestete Niveau wieder.

Tabelle 4: Prozentuale Übereinstimmung der Niveaustufenzuweisungen mit den intendierten Testniveaus in den deutschen Lernerproduktionen

	A1 (*n* = 229)	**A2** (*n* = 228)	**B1** (*n* = 223)	**B2** (*n* = 225)	**C1** (*n* = 225)
Allg. sprachliches Spektrum	24,5%	31,9%	32,5%	28,9%	34,5%
Wortschatz-spektrum	20,5%	34,2%	70,1%	78,2%	59,7%
Wortschatz-beherrschung	31,9%	42,5%	57,6%	54,2%	32,3%
Grammatische Korrektheit	33,2%	45,6%	46,8%	48,0%	19,0%
Kohärenz und Kohäsion	29,7%	35,5%	61,5%	58,7%	35,4%
Sozioling. Angemessenheit	21,0%	42,5%	48,5%	63,6%	27,9%
Orthographische Korrektheit	7,9%	20,2%	46,3%	47,1%	37,6%

Tabelle 4 ist zu entnehmen, dass für den A1-Test "Start Deutsch 1" die Mehrheit der Niveaustufenzuweisungen nicht auf dem intendierten Niveau abgegeben wurden. Im Falle des Bewertungskriteriums "Orthographische Korrektheit" liegen, wie eine genauere Betrachtung zeigt, immerhin 7,8 % der Bewertungen auf den Niveaus B2 bis C2. Für den A2-Test "Start Deutsch 2" verändert sich das Bild nur leicht. Bei fünf der sieben Bewertungskriterien sind Bewertungen auf dem B1-Niveau am häufigsten. Beinahe ein Viertel der Bewertungen liegt für das Kriterium "Orthographische Korrektheit" sogar zwei Stufen und mehr über dem erwartbaren Niveau. Anders als bei den beiden A-Niveaus entsprechen die Bewertungen für das "Zertifikat Deutsch" mit dem Zielniveau B1 weitgehend dem Niveau des Tests. Auffällig sind allerdings auch hier zahlreiche Einstufungen auf den Niveaus B2 und C1 für das Kriterium "Orthographische Korrektheit". Für den Test telc Deutsch B2 liegt die Mehrzahl der Bewertungen tatsächlich auf dem B2-Niveau. Für das Kriterium "Orthographische Korrektheit" sind mit insgesamt 39,1 % der Urteile auf den Niveaus C1 und C2 ungewöhnlich viele Bewertungen auf einem sehr hohen Niveau. Bemerkenswerterweise ist eine substanzielle Menge an Beurteilungen nicht mehr auf dem B2-Niveau. Dieser Trend bestätigt sich bei

dem Test telc Deutsch C1, bei dem die meisten Bewertungen für fünf der sieben Bewertungskriterien eine Stufe unter dem Zielniveau des Tests liegen.

Die bei den deutschen Produktionen beobachteten teilweise erwartungswidrigen Verteilungsmuster finden sich, wenngleich in etwas anderer Form, auch in den Italienisch-Daten, wie Tabelle 5 illustriert:

Tabelle 5: Prozentuale Übereinstimmung der Niveaustufenzuweisungen mit den intendierten Testniveaus in den deutschen Lernerproduktionen

	A1 (n = 229)	**A2** (n = 229)	**B1** (n = 228)	**B2** (n = 218)
Allg. sprachliches Spektrum	15,7%	31,9%	32,5%	28,9%
Wortschatz-spektrum	10%	41%	82,9%	75,2%
Wortschatz-beherrschung	16,6%	41%	67,1%	61%
Grammatische Korrektheit	20,1%	52%	63,6%	32,1%
Kohärenz und Kohäsion	25,8%	59,4%	60,1%	21,1%
Sozioling. Angemessenheit	17%	73,4%	71,5%	28,9%
Orthographische Korrektheit	5,2%	29,3%	10,1%	75,2%

Eine zweite Quelle mit Informationen zur Funktionalität der Bewertungsskalen ergibt sich aus der Berechnung von Korrelationen zwischen den Kriterien (Spearmans *rho*). Insofern die betrachteten GeR-Skalen unterschiedliche Facetten eines gemeinsamen Konstrukts abbilden, sollten sich jeweils mittlere Korrelationen ergeben. Niedrige Korrelationen würden dafür sprechen, dass kein oder ein nur geringer Zusammenhang zwischen den Skalen besteht. Hohe Korrelationen würden im Umkehrschluss darauf hindeuten, dass die Skalen mehr oder weniger austauschbar sind. Die meisten der auf die deutschen Produktionen angewendeten Kriterien korrelieren auf einem mittleren Niveau miteinander (zwischen $r = 0{,}414$ und $r = 0{,}835$, $p < 0{,}05$). Auffällig ist zudem, dass das globale Kriterium "Allgemeines sprachliches Spektrum" tendenziell höher mit den Detailkriterien korreliert als diese untereinander. Dieser Befund deutet darauf hin, dass die Detailkriterien tatsächlich verschiedene Facetten eines gemeinsamen Konstrukts darstellen. Etwas aus dem Rahmen fällt auf dem A1-Niveau das Kriterium "Wortschatzbeherrschung", das nur gering mit dem globalen Kriterium "Allgemeines sprachliches Spektrum"

zusammenhängt (r = 0,296, p < 0,05). Auf dem B2-Niveau korreliert das Kriterium "Orthographische Korrektheit" gering mit den übrigen Kriterien (die Werte liegen zwischen r = 0,233 und r = 0,352, p < 0,05).

Das Muster der Korrelationen in den italienischen Produktionen entspricht weitgehend dem der deutschen Daten. Wiederum auffällig sind hier die geringen Korrelationen der Skala "Orthographische Korrektheit" mit den übrigen Kriterien auf dem B1-Niveau (zwischen r = 0,311 und r = 0,477, p < 0,05) und dem B2-Niveau (r = 0,172, p < 0,05 für die Korrelation mit der Skala "Allgemeines sprachliches Spektrum"; die übrigen Korrelationen waren nicht signifikant).

Als dritte und aufschlussreichste Datenquelle wurde in der vorliegenden Studie für die Deutsch- wie auch für die Italienisch-Daten jeweils eine Multifacetten-Rasch-Analyse durchgeführt. Abbildung 1 gibt für die deutschen Produktionen mit dem Facettenraum eine Übersicht über die drei ausgewerteten Facetten "Kompetenzausprägung der Testteilnehmer", "Strenge der Beurteiler" sowie "Schwierigkeit des Bewertungskriteriums" wieder. Abbildung 2 enthält wiederum eine Übersicht der Facetten aus der Multifacetten-Rasch-Analyse für die Italienisch-Daten.

Die erste Spalte von Abbildung 1 "Maß" gibt die Metrik wieder, auf der alle berücksichtigten Facetten der Testsituation skaliert sind: Logits. Die zweite Spalte "Testkandidat" steht für die Verteilung der Kompetenzausprägung bei den Testteilnehmenden. Es wird deutlich, dass die sprachliche Kompetenz bei der Mehrzahl der Testkandidaten im mittleren Bereich (nahe 0) angesiedelt ist. Demgegenüber gibt es vergleichsweise wenige Testteilnehmende mit sehr geringer Kompetenzausprägung (im deutlich negativen Bereich) und wenige Testteilnehmende mit sehr hoher Kompetenzausprägung (im deutlich positiven Bereich). Die Spalte "Bew." bildet die Strenge bzw. Milde der Bewertenden A und B ab. Wie aus Abbildung 1 hervorgeht, unterscheiden sich die beiden Bewertenden kaum in ihrer Strenge. Aus Spalte 4 wird die Schwierigkeit der verwendeten Bewertungskriterien ersichtlich. Das schwierigste Kriterium stellt "Grammatische Korrektheit" mit einem Logit-Wert von 1,0 dar, während das Kriterium "Orthographische Korrektheit" mit -1,36 Logits mit Abstand am leichtesten ist. Im mittleren Bereich befinden sich die Testkriterien "Wortschatzspektrum" (-0,37 Logits), "Allgemeines sprachliches Spektrum" (-0,03 Logits), "Soziolinguistische Angemessenheit" (0,12 Logits), "Kohäsion und Kohärenz" (0,29 Logits) sowie "Wortschatzbeherrschung" (0,34 Logits). Damit liegt zwischen dem leichtesten und dem schwersten Kriterium eine Spanne von fast 2,37 Logits, was bezogen auf die Referenzrahmenskala einen Unterschied von einer kompletten Niveaustufe ausmachen kann. Die Spalte "Skala" drückt aus, mit welchem Kompetenzgrad ein Testteilnehmender ein bestimmtes Referenzniveau zugesprochen bekommt. Auffällig ist hier, dass die Abstände zwischen den Referenzniveaus ungleich sind.

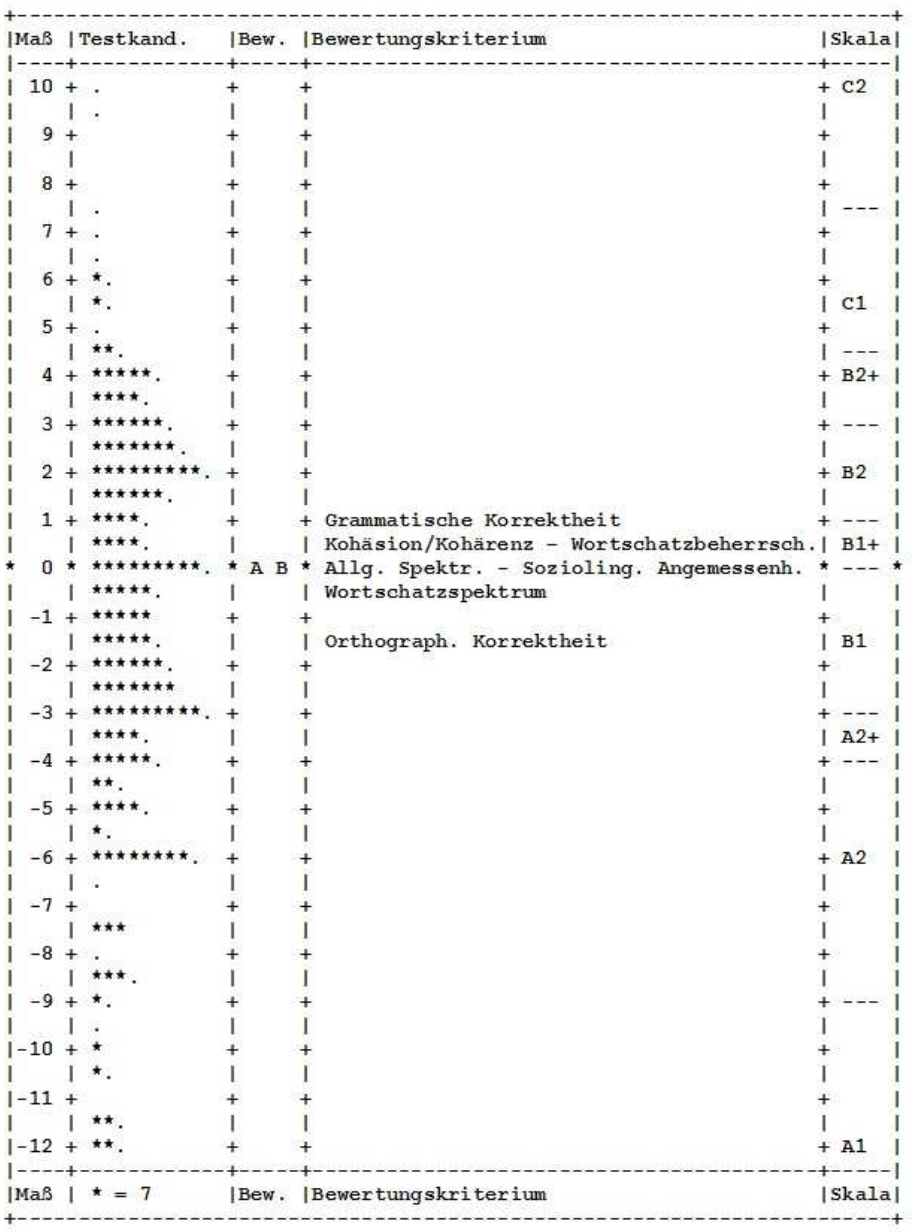

Abbildung 1: Facettenraum für die Deutsch-Produktionen

Bemerkung: In Spalte 2 "Testkandidat" werden jeweils sieben Testkandidaten durch einen Stern repräsentiert.

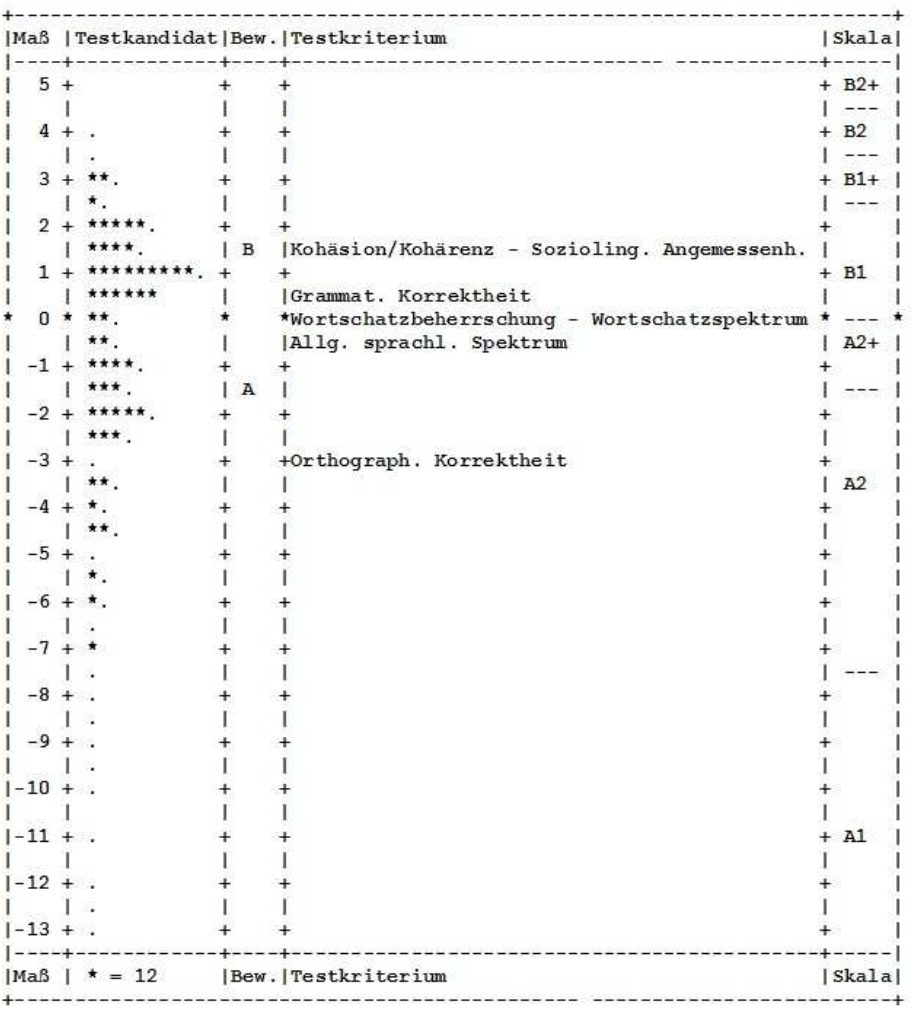

Abbildung 2: Facettenraum für die Italienisch-Produktionen

Bemerkung: In Spalte 2 "Testkandidat" werden jeweils zwölf Testkandidaten durch einen Stern repräsentiert.

Eine detailliertere Betrachtung der einzelnen Facetten bestätigt diese Befunde. Insbesondere zeigt sich bei der Fit-Statistik für die Facette "Testkriterium", dass das Kriterium "Orthographische Korrektheit" mit einer mittleren quadratischen Abweichung von 2,06 beim Infit deutlich mehr Varianz produziert, als vom Rasch-Modell vorhergesagt wird. In der Literatur werden mittlere quadratische Abweichungen von mehr als 1,4 als problematisch angesehen (Bond & Fox 2007).

Der Infit-Wert für "Orthographische Korrektheit" stellt jedoch insofern eine Ausnahme dar, als der über alle Testkriterien aggregierte Infit-Wert 1,04 beträgt und damit nahe am Idealwert von 1,0 liegt. Auch die Separationsreliabilität für die Facette "Testkriterium" von 1,0 deutet für alle Skalen zusammen darauf hin, dass diese zwischen Testteilnehmenden unterschiedlicher Kompetenzstufen sicher differenzieren können.

Die Verteilung der Kompetenzausprägungen bei den Testteilnehmenden in Abbildung 2 ist vergleichbar mit der in den Deutschtests. Anders als dort sind die beiden Bewertenden jedoch deutlich unterschiedlich streng. Die Differenz zwischen dem milden Bewertenden A und dem strengen Bewertenden B beträgt, wie die Multifacetten-Rasch-Analyse ergab, 3,04 Logits. Dies kann zu Unterschieden bei den endgültigen Referenzrahmeneinstufungen von bis zu zwei Niveaustufen führen. Mit Blick auf die Schwierigkeit der Bewertungskriterien ergibt sich ebenfalls ein auffälliger Befund: Wie bereits in den Deutsch-Daten streut die Schwierigkeit breit. Die Differenz zwischen dem leichtesten Kriterium – wiederum "Orthographische Korrektheit" (-2,94 Logits) – und dem schwersten "Kohäsion und Kohärenz" (1,69 Logits) beträgt 4,62 Logits und damit bis zu zwei volle Referenzrahmenstufen. Die Bewertungskriterien "Allgemeines sprachliches Spektrum" (-0,31), "Wortschatzspektrum" (-0,13 Logits), "Wortschatzbeherrschung" (-0,11 Logits), "Grammatische Korrektheit" (0,51 Logits) und "Soziolinguistische Angemessenheit" (1,31 Logits) liegen zwischen diesen beiden extremen Bewertungskriterien. Hinsichtlich der Skalenstruktur fallen noch ausgeprägter als bei den Deutsch-Daten die ungleichen Abstände zwischen den Referenzstufen auf. Mit Blick auf die Fit-Statistik für die Facette "Testkriterium" sticht "Soziolinguistische Angemessenheit" hervor, die mit einer mittleren quadratischen Abweichung von 1,44 über dem als kritisch angesehenen Schwellenwert liegt. Insgesamt liegt der Fit-Wert für alle Testkriterien zusammen jedoch bei einer mittleren quadratischen Abweichung von 0,95 und damit nahe beim Idealwert. Die Separationsreliabilität der Testkriterien ist mit 1,0 wiederum sehr stark ausgeprägt, d. h. die Testkriterien unterscheiden sich deutlich voneinander.

4. Diskussion und Ausblick

Ziel der vorliegenden Studie war es, die Funktionalität ausgewählter Referenzrahmenskalen vor dem Hintergrund authentischer Lernersprache zu überprüfen. Die Bilanz fällt diesbezüglich gemischt aus. Einerseits zeigen die Korrelationen zwischen den GeR-Skalen mittlere Zusammenhänge. Dies lässt sich so interpretieren, dass die verschiedenen Bewertungskriterien unterschiedliche Aspekte ein

und desselben Konstrukts abbilden. Auch korrelieren die Detailskalen in der Regel stärker mit der Globalskala "Allgemeines sprachliches Spektrum" als untereinander. Dies unterstützt die vorangegangene Interpretation nochmals. Allerdings zeigte sich auch, dass das Bewertungskriterium "Orthographische Korrektheit" teilweise nur wenig mit den anderen Bewertungskriterien korreliert. Dies stellt ein erstes Indiz dafür dar, dass das Kriterium "Orthographische Korrektheit" mit Blick auf die Skalenfunktionalität aus dem Rahmen fällt. Ob es sich hierbei um einen Skaleneffekt oder aber möglicherweise um einen Bewertereffekt handelt – etwa um ein subjektives Bewertungskonstrukt – kann allerdings eindeutig nur im Rahmen weiterer Untersuchungen mithilfe qualitativer Daten aufgeklärt werden.

Weitere Hinweise auf die Funktionalität der übrigen Skalen stellen die Fit-Statistiken aus den Rasch-Analysen dar, die mit Ausnahme der Kriterien "Orthographische Korrektheit" und "Soziolinguistische Angemessenheit" relativ gut den Modellerwartungen entsprechen. Mit anderen Worten: Die untersuchten Skalen verhalten sich überwiegend konform mit den Erwartungen des statistischen Modells, was sich als Beleg für ihre Validität interpretieren lässt.

Diese Schlussfolgerung wird jedoch durch die kontraintuitiven Häufigkeitsverteilungen der Niveaustufenurteile konterkariert. Ein substanzieller Anteil der zugewiesenen Niveaustufen liegt bei beiden betrachteten Sprachen auf den unteren Niveaus über dem intendierten Niveau der Tests. Umgekehrt – wenn auch weniger stark ausgeprägt – liegen die Niveaustufenzuweisungen bei den oberen Niveaus in vielen Fällen unter den intendierten Testniveaus. Zwar könnten die beobachteten Effekte theoretisch auch auf eine Zentraltendenz der Beurteilenden zurückgehen. Die Infit-Statistiken für die Bewerter-Facette liefern hierfür jedoch keine Hinweise (der Infit beträgt für den deutschen Rater A 0,98 und für Rater B 1,30, für den italienischen Rater A 0,93 und für Rater B 1,18). Damit sind die mithilfe der GeR-Skalen durchgeführten Bewertungen nicht in vollem Umfang erwartungstreu.

Die ausgeprägte Streuung bei der Schwierigkeit der Bewertungskriterien deutet überdies darauf hin, dass diese zumindest nicht uneingeschränkt bei der Bewertung von Lernenden herangezogen werden können. Auch wenn nur die deutschen Skalen von den Verfassern des GeR empirisch überprüft worden waren, so ist es für die Validität der Kriterien allemal problematisch, wenn verschiedene Skalen, die vom theoretischen Anspruch her demselben GeR-Niveau zuzuordnen sind, jedoch faktisch unterschiedlich schwierig sind. Einer inhaltlichen Klärung bedarf in jedem Fall, warum das Kriterium "Orthographische Korrektheit" bei beiden Sprachen so vergleichsweise leicht ist. Ermöglichen es vielleicht spezifische Strategien der Testteilnehmer, hier besonders gut abzuschneiden? Liegt es an linguistischen Spezifika der Zielsprachen wie etwa einer geringen orthographischen Tiefe? Oder bedarf die Formulierung der Kriterien ggf. einer Revision?

Und warum ergibt sich in jeder Sprache eine andere Schwierigkeitshierarchie der Kriterien? Fragen wie diese können freilich nur mithilfe weiterer Datenquellen und damit neuer Studien beantwortet werden.

Auch die Bewerter-Facette wirft Fragen auf. Wenn auch die beiden Bewertenden der Deutsch-Tests sich in ihrer Strenge kaum unterschieden, so sind die Strengeunterschiede bei den Bewertenden der Italienisch-Tests deutlich. Mit anderen Worten: Offensichtlich wenden die beiden Bewertenden die GeR-Skalen (trotz der vorangegangenen intensiven Schulung mit dem Ziel, Strengeunterschiede möglichst zu nivellieren) auf unterschiedliche Weise an. In der Folge wird derselbe Lernende in Abhängigkeit des Bewertenden jeweils einem anderen Niveau zugewiesen, was der dem GeR immanenten Idee der Standardisierung erkennbar zuwider läuft. Zwar ist das beschriebene Problem auch aus vielen anderen Bewertungskontexten bekannt und lässt sich im Rahmen von Multifacetten-Raschanalysen – ein insgesamt konsistentes Bewerterverhalten vorausgesetzt – durch die Berechnung eines sog. *Fair Average* beheben. Für einen standardmäßigen Einsatz in der Praxis erscheint dieses Verfahren jedoch zu voraussetzungsreich und aufwändig. Außerdem drängt sich die Schlussfolgerung auf, dass die Anwendung von Skalensystemen gleich welcher Art intensiv von den Bewertenden trainiert werden muss. Ein Monitoring der Bewertenden muss dann zeigen, ob ein zunehmend übereinstimmendes Bewerterverhalten erreicht wird. Falls dies nicht der Fall sein sollte, müssten die Bewertenden bzw. die Skalen einer weiter gehenden kritischen Untersuchung unterzogen werden.

In der vorliegenden Studie wurden ausgewählte GeR-Skalen auf authentische schriftliche Produktionen von Lernenden angewendet. Hierbei zeigte sich insgesamt, dass die auf der Basis von subjektiven Expertenurteilen konstruierten GeR-Skalen in der Praxis zumindest nicht unproblematisch eingesetzt werden können, da die Bewertungskriterien klar unterschiedlich schwer sind, da teilweise kontraintuitive Niveauzuschreibungen abgegeben werden, und weil ein statistischer Zusammenhang zwischen den Skalen nicht durchgängig gegeben ist. Eine solchermaßen beeinträchtigte Funktionalität schränkt die Nutzung des GeR in praktischen Zusammenhängen selbstverständlich erheblich ein und steht auch in einem Spannungsfeld zum Anspruch des GeR, Lernersprache mit einem einheitlichen deskriptiven Raster zu beschreiben.[4]

Die vorliegende Untersuchung ist allerdings insofern von begrenzter Reichweite, als sie lediglich Indizien dafür zusammengetragen konnte, dass es bei einer Anwendung von GeR-Skalen auf authentische Lernersprache zu Auffälligkeiten kommt, dass sich diese bei der Anwendung auf authentische Lernersprache also

4 Nicht zuletzt aus diesem Grund haben Harsch & Martin (2012) auf die Notwendigkeit hingewiesen, GeR-Skalen in praktischen Zusammenhängen an die jeweiligen Spezifika der Bewertungssituation anzupassen.

teilweise nicht bewähren. Wie diese Auffälligkeiten zustande kommen, darüber kann die Studie bedingt durch das gewählte Forschungsdesign keine Aussagen machen. Überdies können die beobachteten problematischen Aspekte der Bewertung trotz der vergleichsweise großen Stichprobe und der hohen Qualität der Bewertungen auch nicht eindeutig auf die Bewertungskriterien zurückgeführt werden. Die beobachteten Effekte können beispielsweise auch aus Spezifika der Lernertexte oder aus Beurteilereffekten resultieren.

Aus der Forschungsperspektive erscheint es daher notwendig aufzuklären, wodurch die beobachteten Auffälligkeiten bei der Anwendung des GeR zustande kamen. Qualitative Studien z. B. mit Laut-Denken-Protokollen könnten helfen zu verstehen, in welcher Weise Bewertende Gebrauch von den GeR-Skalen machen bzw. welche Ursachen spezifisch Bewerterprobleme haben; Fokusgruppen könnten dabei helfen, die vorhandenen Skalen so zu adaptieren, dass plausiblere Ergebnisse resultieren. Schließlich könnten auch korpuslinguistische Untersuchungen von Lernersprache sprachspezifisch nachzeichnen, wieso bestimmte Skalen leichter bzw. schwerer sind als andere. Korpuslinguistische Untersuchungen von Lernersprache wären zudem insofern nötig, die vorhandenen GeR-Skalen empirisch zu überprüfen. Möglicherweise würde sich dabei ergeben, dass bestimmte Skalen Lernersprache bzw. ihre Entwicklung nur ungenügend erfassen. Alle die genannten Forschungen würden in jedem Falle dazu beitragen, die, wie aus der Literatur hervorging, geringe theoretische Anbindung und empirische Überprüfung der GeR-Skalen zu verbessern, diese also zu validieren.

Nur vor dem Hintergrund solcher Forschungen ist es letzten Endes möglich, die GeR-Skalen als robuste und aussagekräftige Indikatoren für fremdsprachliche Kompetenz zu verwenden. Angesichts der vielen weitreichenden Entscheidungen, die auf der Grundlage von GeR-Bewertungen getroffen werden – etwa über Schulerfolg, die Zulassung zum Studium, die Auswahl für eine Stelle, eine Beförderung – erscheint eine gründliche weitergehende Validierung der GeR-Skalen dringend geboten.

Eingang des revidierten Manuskripts 08.01.2016

Literaturverzeichnis

Alderson, Charles C. (2007), The CEFR and the need for more research. *Modern Language Journal* 91: 4, 659-663.

Bond, Trevor; Fox, Christine (2007), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Boyd, Adriane; Hana, Jirka; Nicolas, Lionel; Meurers, Detmar; Wisniewski, Katrin; Abel, Andrea; Schöne, Karin; Štindlová, Barbora; Vettori, Chiara (2014), *The MERLIN corpus: Learner language and the CEFR*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, May 26-31, 2014.

Barkowski, Hans (2003), Skalierte Vagheit – der europäische Referenzrahmen für Sprachen und sein Versuch, die sprachliche Kommunikationskompetenz des Menschen für Anliegen des Fremdsprachenunterrichts niveaugerecht zu portionieren. In: Bausch, Karl Richard; Christ, Herbert; Königs, Frank G., Krumm, Hans-J. (Hrsg.), 22-28.

Bausch, Karl-Richard; Christ, Herbert; Königs, Frank G., Krumm, Hans-J. (Hrsg.), *Der Gemeinsame Europäische Referenzrahmen für Sprachen in der Diskussion. Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts*. Tübingen: Narr.

Cambridge ESOL (2011), *Using the CEFR: Principles of Good Practices*. Cambridge: Cambridge ESOL.

Cools, Dorien & Sercu, Lies (2006), Die Beurteilung von Lehrwerken an Hand des Gemeinsamen Europäischen Referenzrahmens für Sprachen: Eine empirische Untersuchung von zwei kürzlich erschienenen Lehrwerken für Deutsch als Fremdsprache. *Zeitschrift für interkulturellen Fremdsprachenunterricht* 11: 3 [Online: https://zif.spz.tu-darmstadt.de/jg-11-3/docs/Cools_Sercu.pdf, 18.12.2015].

Eckes, Thomas (2010), Facetten der Genauigkeit. Zur Reliabilität der Beurteilung fremdsprachlicher Leistungen. *Deutsch als Fremdsprache* 48: 4, 195-204.

Eckes, Thomas (2015), *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2. rev. und erw. Aufl.). Frankfurt am Main: Lang.

Europäische Kommission (2013), *Die Umsetzung des Gemeinsamen europäischen Referenzrahmens für Sprachen in den europäischen Bildungssystemen* [Online: http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/IPOL-CULT_ET%282013%29495871_DE.pdf, 31.08.2015].

Europarat (2001), *Gemeinsamer europäischer Referenzrahmen für Sprachen. Lernen, Lehren, Beurteilen*. Berlin, München: Langenscheidt.

Figueras, Neus (2012), The impact of the CEFR. *ELT Journal* 66: 4, 477-485.

Harsch, Claudia (2005), *Der Gemeinsame Europäische Referenzrahmen für Sprachen: Leistung und Grenzen. Die Bedeutung des Referenzrahmen im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt* [Online: http://opus.bibliothek.uni-augsburg.de/opus4/files/297/DISS_Claudia_Harsch.pdf, 03.09.2015].

Harsch, Claudia & Martin, Guido (2012), Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17: 4, 228-250.

Hilpisch, Kai (2012), *Gemeinsamer Europäischer Referenzrahmen für Sprachen: Der GeR im Überblick*. Hamburg: Diplomica.

House, Juliane (2003), Der Gemeinsame europäische Referenzrahmen für Sprachen – Anspruch und Realität. In: Bausch, Karl-Richard; Christ, Herbert; Königs, Frank G., Krumm, Hans-J. (Hrsg.), 95-105.

Hulstijn, Jan H. (2007), The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91, 663-667.

Kusseling, Françoise & Lonsdale, Deryle (2013), A corpus-based assessment of French CEFR lexical content. *Canadian Modern Language Review* 69: 4, 436-461.

Linacre, Michael (2009), FACETS Rasch measurement computer program. Chicago: MESA Press.

Quetz, Jürgen (2003), Der Gemeinsame Europäische Referenzrahmen: Ein Schatzkästlein mit Perlen, aber auch mit Kreuzen und Ketten... In: Bausch, Karl-Richard; Christ, Herbert; Königs, Frank G., Krumm, Hans-J. (Hrsg.), 145-155.

Quetz, Jürgen (2007), Standards und Kompetenzentwicklung in Fremd- und Zweitsprachen: Der Gemeinsame europäische Referenzrahmen für Sprachen und das Europäische Sprachenportfolio. In Reich, Hans H.; Roth, Hans-J.; Neumann, Ursula (Hrsg.), *Sprachdiagnostik im Lernprozess. Verfahren zur Analyse von Sprachständen im Kontext von Zweisprachigkeit*. Münster et al.: Waxmann, 43-54.

Roche, Jörg (2005), Von der Spracherwerbsforschung zur Diagnostik und Standardentwicklung. In: Bausch, Karl-Richard; Burwitz-Melzer, Eva; Königs, Frank G.; Krumm, Hans-Jürgen (Hrsg.), *Bildungsstandards für den Fremdsprachenunterricht auf dem Prüfstand. Arbeitspapiere der 25. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Narr, 227-239.

Schneider, G. (2005), Der ‚Gemeinsame europäische Referenzrahmen für Sprachen' als Grundlage von Bildungsstandards für die Fremdsprachen – Methodologische Probleme der Entwicklung und Adaptierung von Kompetenzbeschreibungen. *Schweizerische Zeitschrift für Bildungswissenschaften* 27: 1, 13-36.

Simons, Mathea & Colpaert, Jozef (2015), Judgmental evaluation oft he CEFR by stakeholders in language testing. *Revista de Lingüística y Lenguas Aplicadas* 10: 2015, 66-77 [Online: http://polipapers.upv.es/index.php/rdlyla/article/view/3434/4081, 18.12.2015].

Vogt, Karin (2011), *Fremdsprachliche Kompetenzprofile: Entwicklung und Abgleichung von GeR-Deskriptoren für das Fremdsprachenlernen mit einer beruflichen Anwendungsorientierung*. Tübingen: Narr Francke Attempto.

Weir, Cyril J. (2005), Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing* 22: 3, 281-300.

Wirtz, Markus & Caspar, Franz (2002), *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wisniewski, Katrin (2014), Die Validität der Skalen des Gemeinsamen europäischen *Referenzrahmens für* Sprachen*. Eine empirische Untersuchung der Flüssigkeits- und Wortschatzskalen des GeRS am Beispiel des Italienischen und Deutschen*. Frankfurt et al.: Lang.