Bildungsstandards für die Erste Fremdsprache: Sprachenpolitik auf unsicherer Basis

Antwort auf das Positionspapier der DGFF

Jürgen Quetz[1] und Karin Vogt[2]

In *ZFF* 2/2008, the *Deutsche Gesellschaft für Fremdsprachenforschung* (*DGFF* = German society for foreign language research) has published a statement in support of the German *Educational Standards for the First Foreign Language*. In this paper we will try to show that there is a major flaw in the underpinnings of these standards which raises the question whether they are an acceptable instrument for the political decisions in the field of standard setting and evaluation at all. Since the standards they set are based on the *Common European Framework's* (CEF) reference levels A2 and B1, we have summarized a number of critical points made with regard to these levels, which all point at major inconsistencies in the system of branching and scaling laid out in the CEF. As an answer to the *DGFF's* statement we hold the view that the use of the CEF's reference levels for standard setting in the German educational system was premature and possibly a most unfortunate mistake.

1. Einleitung

Die DGFF hat in ZFF 2/2008 ein Positionspapier zu den "Bildungsstandards für die Erste Fremdsprache" (im Folgenden kurz "Bildungsstandards") veröffentlicht, das im Tenor sehr positiv ist und nur einige wenige Hinweise auf Schwächen dieses Dokuments enthält, die wiederum eher auf die Frage der Inhalte und der Modellierung von "Bildung" im Kompetenzbegriff nach Weinert (2001) abzielen. Wir schließen uns der Meinung der DGFF an, dass Standards für Schulabschlüsse notwendig sind. Da sie aber Lebenschancen beeinflussen, müssen sie so gut und zuverlässig sein wie nur irgend möglich. Standards auf unsicherer Basis sind eher schädlich, da sie Objektivität vorgaukeln, wo diese nicht gegeben ist. Unser Beitrag bezieht sich daher allein auf die Frage: Wie überzeugend und valide ist das Konstrukt, das den Niveaus A2 bzw. B1 zugrunde liegt? Im Positionspapier finden sich Ausführungen hierzu in den Abschnitten 3.2 bis 3.4 zu Kompetenzstrukturmodellen, Kompetenzstufen und *standard setting*. Dieses ist der einzige Teil des Papiers, zu dem wir Stellung nehmen, aber es ist zugleich der zentrale, denn mit den dort positiv rezipierten Verfahren steht und fällt der ganze Versuch

1 Seniorprof. Dr. Jürgen Quetz, Leuphana Universität Lüneburg, Scharnhorststr. 1, 21335 Lüneburg, E-Mail: quetz@leuphana.de
2 Prof. Dr. Karin Vogt, Pädagogische Hochschule Heidelberg, Im Neuenheimer Feld 561, 69120 Heidelberg, E-Mail: vogt@ph-heidelberg.de

der Bildungsstandards, Kompetenzniveaus für Schulabschlüsse festzuschreiben. Hier weicht unsere Einschätzung von der der *DGFF* ab.

Die Deskriptoren, mit deren Hilfe die Bildungsstandards den erwünschten *output* beschreiben, sind mehr oder minder direkt aus dem "Gemeinsamen europäischen Referenzrahmen" (Council of Europe / Europarat 2001, hinfort kurz GeR 2001) übernommen. Notwendig ist deshalb der Blick auf dieses Dokument und seine Tragfähigkeit als Fundament weitreichender bildungspolitischer Entscheidungen. Obgleich unser Beitrag somit vor allem auf den GeR bezogen ist, befasst er sich ganz zentral mit den Bildungsstandards, deren momentane Formulierung wir gerade in diesem Punkt für revisionsbedürftig halten.

Unser Beitrag geht davon aus, dass die Bildungsstandards in erster Linie *Sprach*standards sind. Wir äußern uns allerdings *passim* auch zu weiteren Problemen, sofern sie für die Diskussion eines Modells von Sprachkompetenz relevant sind. Dazu gehört u.a., welche Domänen z.B. in seinen Deskriptoren berücksichtigt sind und welches Konzept von soziolinguistischer bzw. interkultureller Kompetenz vertreten wird. Im Zentrum stehen aber Fragen der theoretischen Basis der Bildungsstandards, und diese Basis für die eingeforderten Kompetenzniveaus stellt der GeR dar. Hier gibt es eine Reihe von offenen Fragen, die bislang in der Fachdiskussion wenig Aufmerksamkeit gefunden haben. Wir sind in erster Linie an einer Überprüfung der Plausibilität des *scaling-and-branching*-Konzepts des GeR interessiert. Eignet sich der GeR überhaupt für die Übernahme in Bildungsstandards? Wenn ja, wie geglückt ist diese Übernahme?

2. Zur Geschichte der Beziehungen zwischen GeR und Bildungsstandards

Der GeR wurde 2001 zum Europäischen Jahr der Sprachen mehr oder minder zeitgleich vom Europarat auf Englisch und vom Goethe-Institut und seinen Partnern auf Deutsch veröffentlicht. Seine Ziele sind vielfältig: Vor allem ist er ein sprachenpolitisches Dokument, das der Förderung der Mehrsprachigkeit in Europa dienen soll. Daneben gilt er als Kompetenzmodell, mit dessen Hilfe man methodische Ansätze, Materialien und Qualifikationen beschreiben und vergleichen kann. Er ist auch als Basis für transparente Tests und zur Vergleichbarkeit von Abschlüssen etc. gedacht. Vor allem diese Komponente führte dazu, dass sich ein einzelner Aspekt des GeR wie ein Buschfeuer verbreitete: das System der sechs Referenzniveaus. Man übersah dabei, dass es wegen des prinzipiell offenen *scaling*-Systems durchaus auch mehr oder weniger Stufen auf der Leiter zum Erfolg (= *mastery*, C2) hätten sein können: "Plus"-Stufen sind ja von A2 bis B2

ausformuliert, sodass wir es im Grunde ohnehin schon mit einem neunstufigen System zu tun haben. Es wird häufig auch übersehen, dass potenziell unendlich viele Zwischenstufen definiert werden können, wie die Autoren des GeR vorschlagen (GeR 2001: 41), um so ein flexibles Verzweigungssystem zu erreichen, das den lokalen Bedürfnissen am besten entspricht. Im Schweizer Forschungsprojekt, in dessen Rahmen die gemeinsame Referenzskala entstanden ist, bestand das Endprodukt aus einer zehnstufigen Skala (Schneider & North 2000). Die Testanbieter in Europa, organisiert in der ALTE (= *Association of Language Testers in Europe*), vor allem die ESOL *Main Suite Examinations*, waren vor der Entstehung des GeR noch an einem fünfstufigen System orientiert.

Das Herzstück des Dokuments, die Skalen und Deskriptoren, wurden wegen ihrer Plausibilität auch für Nutzer, die den Inhalt des Dokuments gar nicht kannten, zum alleinigen Referenzpunkt. Dementsprechend unkritisch erfolgte auf Seiten von Bildungspolitik, Lehrkräften, Lernenden, Verlagen und bisweilen auch der Fachwelt die Bezugnahme auf die sechs Niveaustufen. Ebenso fanden sie 2003 bzw. 2004 Eingang in die Bildungsstandards (vgl. Klieme et al. 2003, KMK 2003). Die ursprüngliche Begeisterung für die Skalen wich aber schnell einer gewissen Ernüchterung, weil sich die kritischen Stimmen mehrten, die nicht nur auf die mangelnde Bodenhaftung im Bereich der berücksichtigten Domänen verwiesen (in Bredella 2003 z.B. auf die Vernachlässigung von Inhalten, die im deutschen Bildungswesen eine traditionell wichtige Rolle spielen), sondern direkt das *scaling-and-branching*-Modell des GeR ins Visier nahmen. Fulcher (2004a) veröffentlichte in *The Guardian* – einer Wochenzeitung! – einen kritischen Artikel unter dem Titel: "*Are Europe's tests being built on an 'unsafe' framework?*" und führte diese Beobachtungen (Fulcher 2004b) in *Language Assessment Quarterly* weiter aus. Zuvor waren auch vom Mitverfasser dieser Stellungnahme immer wieder Zweifel an der Plausibilität der Skalen des GeR angemeldet worden (Quetz 2002, 2003, 2004).[3]

Besonders irritierend war aber ein Bericht einer Gruppe führender europäischer Testexperten über den Versuch, einen Sprachtest mit Hilfe des GeR zu validieren (Alderson et al. 2004, vgl. auch Alderson et al. 2006). In diesem Dokument sind fast alle Probleme des GeR-Modells von Skalierungen von Sprachkompetenz beleuchtet; wir gehen im Folgenden noch genauer darauf ein.

An dieser Stelle kann festgehalten werden: Der GeR, gedacht als sprachenpolitisches Leitdokument ohne normative Ansprüche, hatte sich mit einer seiner Komponenten verselbständigt, und die Niveaustufenbezeichnungen A1 bis

3 Zur Klarstellung sei erwähnt, dass der Mitverfasser dieses Beitrags zwar die deutsche Übersetzung (mit)erstellt hat, dabei aber die im Schweizer Forschungsprojekt (s. u.) entstandenen deutschen Skalen nicht verändern durfte. Die Deskriptoren liegen übrigens in Deutsch, Englisch und Französisch vor (Schneider & North 2000).

C2 existieren heute als oft unreflektierte Bezeichnung, deren theoretische Grundlagen vielen Nutzern unklar zu sein scheinen. Sie wurden aber trotzdem als Basis für Objektivierungen im Bildungssystem gewählt. Man kann sagen, dass die Referenzniveaus, ihrer theoretischen Basis beraubt, zu einer Art *travelling concept* verkommen sind. Sie bedeuten heute nicht mehr als die Alltagstheorien hinter den angelsächsischen Begriffen *beginners, intermediate* und *advanced*, im GeR-Jargon "A, B, C", aufgeteilt in jeweils zwei weitere Stufen "1 und 2". Für die allermeisten Nutzer scheint diese Reduktion völlig hinreichend zu sein, und die Siglen werden mittlerweile für Materialien aller Art vergeben, wobei der Bezug zum GeR in aller Regel eher behauptet als nachgewiesen wird. Wie kompliziert so ein Nachweis des Bezugs auf den GeR zu führen ist, hat sich bei der Erstellung eines Handbuchs gezeigt, in dem das Verfahren erläutert wird, wie man Sprachtests auf GeR-Niveaus beziehen kann; daran arbeiten führende europäische Testexperten nun schon jahrelang (Council of Europe 2003, 2009).

Die vorschnelle Übernahme der Referenzniveaus des GeR in die Bildungsstandards hat ihren Teil zur Misere der wuchernden Adaptationen beigetragen, und auch auf dem Weltmarkt der Sprachen ist mittlerweile die gleiche Unbefangenheit zu beobachten. Im Konkurrenzkampf um Marktanteile hat sogar das TOEFL/TOEIC-System seine Tests auf den GeR bezogen, weil das augenscheinlich plausible und übersichtliche GeR-System den britischen Testsystemen mittlerweile einen wichtigen Marktvorteil einzubringen scheint.

Als die KMK also den GeR adoptierte, war die wissenschaftliche Diskussion erst in ihren Anfängen. Hier sind möglicherweise noch Nachbesserungen zu erwarten, die die Kritik am GeR berücksichtigen und umsetzen.

Im Folgenden werden wir einige Probleme mit dem GeR umreißen, wobei wir natürlich die Leistung der Autoren anerkennen. Unternehmen wie dieses sind eine Herausforderung, bei denen oft Schritte gewagt werden müssen, deren Diskussion viel Zeit erfordert. Die hier von uns besonders genau beleuchteten Skalen sind im Rahmen eines Schweizer Forschungsprojekts in akribischer Arbeit, aber unter Zeitdruck erstellt worden (Schneider & North 2000). In der Erprobungsfassung des GeR, die ab 1997 in Expertenkreisen diskutiert wurde, waren sie Teil des Anhangs; das vollständige System war damals noch in der Entwicklung, musste aber bis zum Europäischen Jahr der Sprachen 2001, der zeitlichen Zielvorstellung des Europarats, abgeschlossen und in das Dokument eingearbeitet sein. Manche Probleme sind also offensichtlich durch den hohen Zeitdruck zu erklären, unter dem das Team arbeitete.

3. Stärken des GeR und der in die Bildungsstandards übernommenen Aspekte

Zunächst ist positiv festzuhalten, dass der GeR in den letzten Jahren wie kaum ein anderes bildungspolitisches Dokument die fremdsprachendidaktische Diskussion angeregt und beeinflusst hat, und zwar sowohl auf nationaler als auch auf internationaler Ebene. Der Erfolg des Dokuments über Europa hinaus ist unbestritten. Die wichtigsten Ziele des Europarats, die damit verbunden sind, sind die Förderung von Mehrsprachigkeit und die erhöhte Transparenz von Prüfungen und Zertifikaten etc. und damit einhergehend eine Förderung der Mobilität innerhalb (und *de facto* auch außerhalb) Europas. Es ist noch nicht abzusehen, ob alle diese Ziele erreicht werden können.

Die Skalen und Deskriptoren des GeR in ihrer Funktion als eine gemeinsame Referenzskala sind als eine neuere Entwicklung zu sehen, die mit den bisher bekannten und benutzten Formaten wie beispielsweise *rating scales* aus Tests und Zertifikaten wenig zu tun hat. Die Skalen und Deskriptoren sind empirisch vergleichsweise gut fundiert und methodologisch am klarsten durch einen Blick auf ihre Entwicklung zu verstehen (vgl. North 2000). Dabei gingen die Verfasser auf der Basis eines im Anschluss an traditionelle Modelle definierten Kompetenzbegriffs für Fremdsprachen und anhand von empirischen Daten vor. Die Skalen und Deskriptoren erheben den Anspruch, dass verschiedene Nutzergruppen in verschiedenen Sprachen mit ihnen arbeiten können. Die spontane Begeisterung über dieses Instrument ist wohl vor allem auf dieses Versprechen zurückzuführen. Heute gibt es bald 40 Übersetzungen des englischen Textes, auch in nichteuropäische Sprachen wie Japanisch und Koreanisch, wobei das im Original aufwändig validierte Deskriptorensystem stets mitübersetzt wurde, was natürlich die empirische Basis der Übersetzungen in Frage stellt. Die besondere Stärke des GeR liegt unseres Erachtens in der Herausforderung, die im Bildungswesen benutzten fachdidaktischen Konzepte zu durchdenken und zu versuchen, Kurse, Materialien und Prüfungen vergleichbar zu machen. Wir sind daher auch nicht prinzipiell gegen ein System von Referenzniveaus eingestellt. Auch die Bildungsstandards sind als ein Schritt in die richtige Richtung zu betrachten; dass sie aber vorschnell den GeR, der ja im Grunde nicht normativ konzipiert ist, als Basis gewählt haben, wird noch zu problematisieren sein.

4. Generelle Probleme und offene Fragen

An dieser Stelle sollen nur schlaglichtartig einige allgemeinere Kritikpunkte resümiert werden, die verdeutlichen, dass der GeR als bildungspolitisches Dokument keineswegs unumstritten ist. Einige davon betreffen mehr oder minder direkt auch die Qualität der Deskriptoren, vor allem, was die Wahl der in ihnen beschriebenen Domänen kommunikativer Aktivitäten anbetrifft.

Bei der Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts 2003 wurde der GeR sehr kritisch diskutiert. Ein Punkt ist die fehlende Erfüllung von Ansprüchen, die die Autorengruppe selbst mit dem Dokument verbunden hatte. Als ein Beispiel sei das Kriterium der Transparenz genannt. Im Referenzrahmen wird gesagt, dass "die Informationen klar und explizit formuliert und für den Benutzer verfügbar und leicht verständlich sein müssen". Dennoch hat der Europarat eine Reihe von *user guides* veröffentlicht, in denen über den GeR hinaus Erläuterungen für bestimmte Zwecke gegeben werden. Auch der wiederholte Hinweis auf methodische Neutralität (z.B. GeR 2001: 140, "[...] ohne Befürwortung eines bestimmten Ansatzes und ohne jeglichen Dogmatismus") wird unterminiert, indem der funktional-notionale Ansatz des Europarats aus den Jahren 1975 bzw. 1990 favorisiert wird, der zudem damals auf das Fremdsprachenlernen Erwachsener abzielte und nicht wie der GeR potenziell alle Arten von Fremdsprachenlernenden, angefangen bei Grundschulkindern, in den Blick nahm. Die Darstellung der Prozesse des Sprachenlernens in Kapitel 6 und das Kapitel 7 zur Methodik des Fremdsprachenlernens sind häufig kritisiert worden, weil sie unzureichend sind; es fehlen beispielsweise psycholinguistische Modellierungen von Produktions- und Rezeptionsprozessen (Tönshoff 2003). Überhaupt wird eine unscharfe Begrifflichkeit bei Aussagen zu Spracherwerb und Sprachvermittlung (Krumm 2003) ebenso bemängelt wie fehlende Definitionen von Begriffen im gesamten Dokument (Kleppin 2003). Zwar ist der GeR ein bildungspolitisches Dokument, das sprachenpolitische Ziele verfolgt; es darf jedoch kritisch angemerkt werden, dass mit dem Ignorieren der Rolle des Englischen als *lingua franca* nicht nur die sprachenpolitische Realität unzureichend rezipiert, sondern gleichzeitig ein sehr reger Forschungszweig unbeachtet gelassen wird (House 2003). Der Schwerpunkt des Dokuments scheint auf dem Testen zu liegen, andererseits vermisst Vollmer (2003) in Kapitel 9 eine kritische Diskussion von Fragen der reliablen Messbarkeit von einzelnen Kompetenzdimensionen. Die Idealisierung des *native speaker* wird zwar aufgegeben (North 2000: 55), jedoch wird sie nicht ersetzt, z.B. durch das Konzept eines *intercultural speaker*, wie es Kramsch (1998) und Byram (1997, 2008) vertreten.

Bredella (2003), Burwitz-Melzer (2005, 2007) und andere haben wiederholt darauf aufmerksam gemacht, dass die pragmatische Ausrichtung des GeR und in

seiner Folge der Bildungsstandards zu einer erheblichen inhaltlichen Verkürzung und Verarmung des schulischen Fremdsprachenunterrichts führen. Im GeR wird "kreativer Sprachgebrauch" nur mit einem bedauernden Achselzucken als "wichtig", aber auf den 350 Druckseiten des Dokuments doch nur einer knappen Seite für wert befunden. Und die im Gefolge der Bildungsstandards entstehenden Vergleichsarbeiten (z.B. des IQ Hessen 2009 für die 6. Klasse Englisch) zeigen, dass diese Befürchtungen nicht unbegründet waren. Die für den Fremdsprachenunterricht nach diesem Konzept akzeptablen pragmatischen Domänen (z.B. Durchsagen am Flughafen verstehen) mögen zwar für die Erwachsenenbildung plausibel sein, sind aber eine Abkehr von dem, was bislang im schulischen Bereich und auch in den meisten Lehrwerken wichtig war.

Wenn die Bildungsstandards ein für den Fremdsprachenunterricht brauchbares Bezugssystem sein sollen, so müssten sie Prozesse zu beschreiben helfen, und nicht nur auf Produkte fokussieren. Dies scheint uns allerdings nicht unbedingt der Fall zu sein. Die Skalen und Deskriptoren im GeR haben in ihrer Funktion als *real-life scales* (Bachman 1990) einen Schwerpunkt auf der Sprachverwendung: sie beschreiben, was ein Lernender im "richtigen Leben" in der Fremdsprache kann. Mit dieser Schwerpunktsetzung, die bisweilen als instrumentalistische Sichtweise von Sprachverwendung kritisiert wird (Krumm 2003, Schmenk 2004), geht einher, dass Sprachen*lernen* weniger berücksichtigt wird als die *Verwendung* von Fremdsprachen in unterschiedlichen Kommunikationssituationen (House 2003). Die Deskriptoren im Referenzrahmen beschreiben Sprachstände, und damit legen sie einen Schwerpunkt auf Sprache als Produkt. Die Prozessdimension des Lernens von Fremdsprachen wird ausgeklammert (Tönshoff 2003). Die Niveaus sind auch nicht äquidistant, wie immer wieder implizit unterstellt wird, so dass sich allein von den Niveaustufen und deren Beschreibungen, wie sie sich in den Deskriptoren finden, keinerlei Hinweise auf die Prozesse des Fremdsprachenlernens ableiten lassen können.

5. Kritische Analyse der Skalen und Deskriptoren im GeR

In weiteren Veröffentlichungen ist schon früh auf den wohl heikelsten Schwachpunkt des GeR hingewiesen worden, das System der Deskriptoren. Einige der folgenden Beobachtungen sind schon in einem Forschungsbericht von Vogt (2007) in der ZFF aufgeführt bzw. von Quetz (2002, 2003, 2004, 2005, 2007) und von Burwitz-Melzer & Quetz (2006) vorgetragen worden. Sie seien hier doch noch einmal zusammengefasst, um zu dokumentieren, dass die DGFF bei ihrer

insgesamt positiven Würdigung der Bildungsstandards eine Reihe von wohlbekannten Problemen ausgeblendet hat.

Wirft man einen genaueren Blick auf die Entstehung der gemeinsamen Referenzskala, so werden die Schwächen der Skalen und Deskriptoren besonders deutlich. Die Deskriptoren, die zu einer einzigen gemeinsamen Referenzskala zusammengefügt wurden, haben ihren Ursprung in 27 Einzelskalen, die sich in ihrer Funktion teilweise erheblich unterscheiden (zu den unterschiedlichen Funktionen von Skalen vgl. Alderson 1991). Teilweise werden Skalen zur Rückmeldung bezogen auf Lernfortschritte (*stages of attainment*) einbezogen, teilweise solche als Basis genommen, die als *analytic rating scales* in Testsituationen verwendet werden. Der Grad der Spezifik der Ursprungsskalen variiert ebenfalls. Harsch (2006, 2007) kritisiert in diesem Zusammenhang, dass die Basis der Beschreibung in den Quellskalen nicht mitbedacht wird und damit schon der grundlegende Ansatz des GeR nicht valide genug ist für bestimmte Zwecke der Leistungsbeurteilung. Da das ursprüngliche Konstrukt der Quellskalen nicht erhellt wird, ergibt sich ein grundlegendes Validierungsproblem der Skalen überhaupt (Harsch 2006, 2007); mehr dazu in Abschnitt 5.3.

Generell ist die Vorgehensweise bei der Erstellung der gemeinsamen Referenzskala nicht lückenlos dargestellt, und daher mangelt es ihr an Transparenz. Als Beispiel sei die Anfangsphase des Schweizer Forschungsprojekts angeführt, in dessen Rahmen die Allgemeinskala erstellt wurde. Hier wurden provisorische Deskriptoren vorläufigen Niveaustufen zugeordnet, aber wie und durch wen? Die Vorgehensweise erschließt sich dem Leser trotz genauer Lektüre aller Begleitliteratur (Schneider & North 2000, North 2000, North & Schneider 1998) nicht vollständig. Andere Fragen betreffen die Serien von Workshops in zwei Etappen während der Hauptuntersuchung: Waren immer die gleichen Informanten dabei, was einen Lerneffekt und damit sich verändernde Ergebnisse bedeuten würde, oder waren es unterschiedliche Gruppen? Nach welchem Muster wurden die neu zu schaffenden Deskriptoren für strategische Kompetenzen geschrieben? Hier bleiben bezüglich der Forschungsmethodologie einige Fragen unbeantwortet.

5.1 Deskriptoren im Bereich rezeptive Fertigkeiten

Eine Eigenschaft des GeR wird im Dokument selbst als gewollt und sinnvoll dargestellt, erweist sich aber bei der Nutzung des GeR als eine Quelle von Unsicherheiten. Die Deskriptoren sind, bildlich gesprochen, komplexe "Moleküle", die aus mehreren "Atomen" bestehen. Unter anderem bedingt durch deren heterogene Herkunft erscheinen sie aber erratisch und ohne jedes erkennbare System über die jeweiligen denkbaren Positionen des Deskriptors verteilt. Als Beispiel sollen die Niveaus B2 bis C2 der Skala "Hörverstehen allgemein" dienen, die wir hier nicht noch einmal abdrucken (GeR 2001: 71f.). In der folgenden Übersicht sieht man, wie die Positionen in den Deskriptoren besetzt sind:

	qualitativ / quantitativ	sprachliche Merkmale	Themen / Situationen	Textsorten / Medien	Bedingungen & Beschränkungen
C2	keinerlei (Schwierigkeiten)	alle Arten (gesprochener Sprache)	-	Medien / live	auch wenn schnell gesprochen wird, wie Muttersprachler dies tun
C1	genug	-	nicht vertraut / abstrakt / komplex	längere Redebeiträge	auch wenn Details bestätigt werden müssen / bei fremdem Akzent
	ein breites Spektrum	idiomatische Wendungen / umgangssprachliche Ausdrucksformen / Registerwechsel	-	-	-
	-	-	-	längere Reden und Gespräche	auch wenn nicht klar strukturiert / Zusammenhänge nicht explizit
B2	-	gesprochene Standardsprache	(weniger) vertraut / privat, gesellschaftlich, beruflich	direkter Kontakt / Medien	Hintergrundgeräusche / unangemessene Diskursstruktur / starke Idiomatik

Zur Klarheit des Systems für potentielle Nutzer trägt dies nicht bei. Wir finden nicht nur in dieser, sondern in allen Skalen des GeR in bunter Mischung

- quantifizierende Merkmale ("keinerlei / genug"),
- qualifizierende Merkmale ("korrekt", "auf einfache Art", "relativ leicht"),
- sprachliche Merkmale ("idiomatische Wendungen", "gesprochene Standardsprache")
- Merkmale von Themen ("vertraut / routinemäßig", "komplex" / "abstrakt" – "konkret")
- Merkmale von Textsorten oder Situationen ("direkter Kontakt / Medien")
- Einschränkungen ("sofern die Gesprächspartner langsam sprechen", "trotz unangemessener Diskursstrukturen") (vgl. Quetz 2004).

Die Problematik der Formulierungen in den Deskriptoren ist mittlerweile für alle "Kompetenzen" diskutiert worden. Alderson et al. (2004) haben in "*The development of specifications for item development and classification within the CEF – Reading and listening – Final report of the Dutch CEF construct project*" ausführlich die rezeptiven Fertigkeiten Hören und Lesen analysiert. Dort finden die Autoren folgende Schwächen, die wir hier der Übersichtlichkeit halber in Stichworten auflisten:

(1) Terminologie: Synonyme oder nicht?

- Hören A2: *understand, get, follow, identify, infer*
- Lesen B1: *understand, locate, scan, identify, combine, extrapolate, recognise*
- Lesen B2: *understand, scan, monitor, obtain, select, evaluate, locate, identify*

(2) Lücken

- Im Text des GeR erwähnte Konzepte werden in den Skalen nicht aufgegriffen
  (zahlreiche Beispiele).
- *tasks* / *activities* sind ungenau definiert: Was bedeutet z.B. zur Orientierung lesen –Durchsagen und Anweisungen verstehen – usw.?

(3) Inkonsistenzen

- Gleiche oder ähnliche Kann-Beschreibungen finden sich auf verschiedenen Niveaus;
- *recognise* wird auf A1, B1 und C1 benutzt, aber nicht auf A2, B2, C2 – warum?
- die o. a. Verben werden offenbar willkürlich auf verschiedenen Niveaus benutzt;
- Wörterbuchgebrauch wird erst auf B2 (und C1) erwähnt, nicht aber darunter oder auf C2;
- Was ist der Unterschied zwischen *specific information* und *specific predictable information* (beides A2)?
- Textsorten werden mal aufgeführt, mal nicht;
- *simple notices* (A1) / *everyday notices* (A2): Wo liegt der Unterschied?

Zudem sollen die Deskriptoren kriteriumsorientiert sein, sollen unabhängig voneinander und mit einer Ja-/Nein-Entscheidung bearbeitbar sein. Auch das Kriterium der Kürze spielt eine Rolle. Im Prozess der Erstellung und Nachbearbeitung der Deskriptoren jedoch wurden viele dieser Kriterien aufgeweicht. North (2000) berichtet beispielsweise aus den Workshops, in denen Lehrkräften vorläufige Deskriptoren zur Evaluation vorgelegt wurden, dass Deskriptoren mit mehr als ca. 25 Wörtern als zu lang zurückgewiesen wurden. Bei der Datenanalyse in der Hauptuntersuchung führten Deskriptoren, die den ursprünglich geplanten Inhaltsbereich "Grad der Unabhängigkeit in der Kommunikation" darstellten (z.B. wie viel Hilfe ist erforderlich?), zu Inkonsistenzen in den Daten. Um das Problem zu lösen, wurden die Deskriptoren als Einschränkungen umformuliert und an bestehende Deskriptoren angehängt. Diese Vorgehensweise wiederum verlängerte einige Deskriptoren merklich, so dass die Qualitätskriterien Kürze und Klarheit (GeR 2001: 39) nicht durchgängig eingelöst werden.

5.2 Deskriptoren für Schreiben und Sprechen (Interaktion)

Ähnliche Untersuchungen wie die von Alderson et al. (2004) zu den rezeptiven Fertigkeiten haben mittlerweile Harsch (2006, 2007) zum Schreiben und Vogt (Bericht über das Forschungsprojekt in der ZFF 1/2007) zum Sprechen vorgelegt. In beiden Arbeiten wird dargestellt, wo der GeR nachgebessert werden müsste.

Harsch (2006, 2007) wendet eine hermeneutische Analyse der Skalen und Deskriptoren für das produktive Schreiben an, wobei sie die gleichen Gesichtspunkte wie Alderson et al. (2004) zugrunde legt. Sie stellt fest, dass in Bezug auf die Strukturierung die Niveaus zwar relativ konsistent voneinander abgesetzt sind. Inkonsistent sind jedoch die Operationen auf den unterschiedlichen Niveaus und die Beschreibung dessen, was auf welchem Niveau verschriftlicht werden kann. Innerhalb der Niveaus finden sich ebenfalls Inkonsistenzen.

Harsch (2006, 2007) findet ähnliche Terminologieprobleme wie Alderson et al. (2004), und auch Lücken sind vorhanden, hier vornehmlich in Bezug auf Genres bzw. Textsorten. Bezüglich der Skalen in Kapitel 5 des GeR zu linguistischen Kompetenzen, die separat von den Skalen in Kapitel 4 (kommunikative Aktivitäten) analysiert wurden, merkt sie an, dass der Gegenstandsbereich jeweils nicht stringent in der Skala umgesetzt wird. Zum Thema Lücken bemerkt sie, dass nähere Angaben zu Themen sowie zu Frequenz und Schwierigkeit des Wortschatzes fehlen, wobei hier aus einem spezifischen Kontext heraus, nämlich dem des schulischen Fremdsprachenlernens, argumentiert wird. Auch grammatische Strukturen sind innerhalb der Niveaus nicht kohärent charakterisiert.

Vogt (erscheint 2010) untersucht im Bereich "mündliche Interaktion" die Strukturiertheit der Skalen (in Bezug auf die Kategorisierung der Merkmale und die Kohärenz der Beschreibung etc.), die Terminologie und deren Eigenheiten bzw. die Konsistenz sowohl in Bezug auf die Deskriptoren als auch auf die Skalen.

Strukturiertheit

Was den Aspekt der Strukturiertheit anbetrifft, geht die Skala "Mündliche Interaktion allgemein" von Situationen aus und erst in den höheren Niveaus von Themen. Der allgemeine Charakter der Skala schlägt sich auch nieder in den sprachlichen Operationen, die recht konsistent das allgemeine Verb "sich verständigen" beinhalten. Der Grad der Hilfestellung nimmt allmählich ab und spielt ab dem Niveau B1 keine Rolle mehr. Die bewältigten Situationen nehmen in Quantität und Qualität zu, während die zur Verfügung stehenden sprachlichen Mittel ausdifferenziert werden. Die Gradierung der mündlichen Sprachfähigkeit "allgemein" scheint sinnvoll, die Formulierung der Situationen, Themen und sprachlichen Mittel ist zunächst einmal gut nachvollziehbar, obgleich der folgende wichtige Aspekt von anderen Modellen der Sprechfertigkeit nicht ganz eingelöst ist.

Ein Charakteristikum des Modells von Bygate (1987, Weiterentwicklung in Luoma 2004) zum Sprechen ist dessen Prozesshaftigkeit und Reziprozität. Der Reziprozität versuchten die Autoren in den Skalen gerecht zu werden, indem sie die Perspektive des Interaktanten als Hörender und Sprechender berücksichtigten. Auffallend ist bei der nachgeordneten Skala "Muttersprachliche Gesprächspartner verstehen", dass der Fokus ausschließlich auf der Rolle des Hörenden liegt. Man hätte eine solche Skala eher im Bereich der rezeptiven Fertigkeiten erwartet. Die dort zu findende Subskala "Gespräche zwischen Muttersprachlern verstehen" wiederum enthält deutliche interaktive Elemente in den Deskriptoren, z.B. im Proviso zu B2: "(…) dürfte aber Schwierigkeiten haben, sich wirklich an Gruppengesprächen mit Muttersprachlern zu *beteiligen,* die ihre Sprache in keiner Weise anpassen" (GeR 2001: 72, Hervorhebung d. Verf.). Somit kann man zu dem Schluss kommen, dass eine saubere Abgrenzung des Gegenstands "mündliche Interaktion" doch nicht durchgehend erfolgt ist.

Terminologie

Die Operationen, die in den Skalen zu finden sind, sind Verben bzw. einfache und komplexere Sprechhandlungen (z.B. "fragen", "erklären" bzw. "sich entschuldigen", "widersprechen") oder es handelt sich um kleinere Szenarien (*cultural scripts*) wie beispielsweise "eine Mahlzeit bestellen", "den Weg erklären" usw. In diese beiden Kategorien fallen die meisten Operationen. Problematischer ist eine weitere Kategorie, die mangels klarer Operationalisierbarkeit linguistisch nicht fassbar ist; es handelt sich dabei um eher alltagssprachliche Ausdrücke, die in der Sprachwissenschaft nicht gebräuchlich sind. Hierzu gehören Formulierungen wie "zurechtkommen", "Situationen bewältigen" (GeR 2001: 83), "mit Situationen umgehen" (ebd.), "Initiativen ergreifen" (GeR 2001: 85), "sich verständigen" (GeR 2001: 79), "Beziehungen aufrecht erhalten" (GeR 2001: 80), "(bei einer Diskussion) mithalten" (GeR 2001: 81, 82). Der vage und unsystematisch selektive Charakter dieser Formulierungen birgt die Gefahr der divergierenden Interpretationen durch verschiedene Nutzer.

Insgesamt differenzieren sich die sprachlichen Operationen, die den Produktionsaspekt der Interaktion beschreiben, auf den höheren Niveaus aus, allerdings ist auch hier keine Systematik zu erkennen, nach der sie dies tun: Die allgemeineren Operatoren werden auch auf den höheren Niveaus beibehalten, es ist keine stringente Entwicklung von den unteren zu den höheren Niveaus zu beobachten.

Die beschriebenen Operationen werden zum einen eingebettet in klassische Diskurstypen wie "Diskussion", "Alltagsgespräch", aber auch in Subkategorien von Diskursen ("Pläne [diskutieren] = Untertyp zu Diskussion", "Anleitungen [geben]", "seine Meinung [äußern]"). Insbesondere bei sprachlichen Operationen, die auch nonverbal sein können, werden sogar konkrete Objekte genannt, beispielsweise "Waren und Dienstleistungen [anbieten]" (GeR 2001: 84), allerdings hauptsächlich auf den niedrigeren Niveaus. Auf höheren Niveaus (ab B2) erhöht sich der Grad der Abstraktion einerseits (z.B. "eine Vorgehensweise [beschreiben]" (GeR 2001: 84), "Gedanken [ausführen]", "eine Position [vertreten]" (GeR 2001: 82), "feinere Bedeutungsnuancen [deutlich machen]" (GeR 2001: 79), "Redebeiträge [strukturieren]" (GeR 2001: 85), "Standpunkte anderer [kommentieren]" (GeR 2001: 83). Hierbei handelt es sich um Versuche, diskursives Verhalten zu erfassen. Eine Systematik im Aufbau ist allerdings nicht auszumachen.

Die Einschränkungen einer Sprachhandlung beziehen sich in den Deskriptoren sowohl auf die Charakterisierung der Textsorte bzw. Diskursform (z.B. "kurze [Gespräche]", "einfaches Alltags[gespräch]", "Routine[gespräch") als auch auf den Inhalt der Interaktion, u.a. in Bezug auf den Abstraktionsgrad ("vertraute [Dinge]",

"Höflichkeitsformen"), aber auch auf Themen und Situationen ("gängige Alltagssituationen: Unterkunft, Reisen, Einkauf, Essen" (GeR 2001: 84), "Fahrkarten kaufen" (GeR 2001: 84), "einfache Routine[aufgaben]" (GeR 2001: 83), "Gewohnheiten / Alltag[sbeschäftigungen]"). Einschränkungen sind also nicht eindeutig zu kategorisieren, weil sie mehrere Aspekte einschließen, nicht nur die Diskursform, innerhalb derer die Sprache verwendet wird, sondern auch die Inhalte und Kontexte. Die in der Interaktion verwendete Sprache ist ebenso in die Kategorie Einschränkungen einzuordnen, z.B. "einfache Gruß- und Abschiedsformeln gebrauchen" (GeR 2001: 81) wird durch das Adjektiv "einfach" eingeschränkt, aber auch durch die konkrete Benennung der verwendeten Ausdrücke. Carrolls (1980) Kategorie der "Unabhängigkeit" spiegelt sich in Form von Provisos, z.B. "sofern die Gesprächspartner deutlich sprechen und stark idiomatischen Sprachgebrauch vermeiden" (GeR 2001: 81). Die Provisos beziehen sich auf die Hilfestellung von Gesprächspartnern in Bezug auf Wiederholung und Umformulierung, deutlicher Aussprache, Gebrauch vereinfachter Sprache, Vermeidung von dialektgefärbter bzw. idiomatischer Sprache und Tempo des Sprachgebrauchs.

Andere Arten von Einschränkungen sind eher implizit, wenn sie in höheren Niveaus als nicht vorhandene Einschränkungen thematisiert werden, z.B. auf B2 "[…], auch wenn es in der Umgebung störende Geräusche gibt". Dies impliziert, dass auf niedrigen Niveaus paralinguistische Faktoren wie Hintergrundgeräusche einen Problemfaktor darstellen. Insgesamt kann man schlussfolgern, dass die Kategorie der Einschränkungen in den Deskriptoren facettenreich ist und in andere Kategorien wie Themen, Situationen und Gegenstände bzw. Inhalte hineinreicht.

Konsistenz

Die Konsistenz der Kategorien in den Deskriptoren, d.h. deren Besetzung innerhalb der Deskriptoren eines Niveaus bzw. einer Skala, bedarf ebenfalls eines kritischen Blicks. Bei der Analyse der Deskriptoren zur mündlichen Interaktion ergeben sich ähnliche Ergebnisse wie bei Alderson et al. (2004), Quetz (2004) sowie Harsch (2006). Die einzelnen Kategorien in den Deskriptoren sind nicht durchgängig besetzt. Bei der Betrachtung der Besetzung der Elemente in den Deskriptoren ergibt sich eine große Spannweite, die sich von einer einzigen besetzen Kategorie ("Kann sich beschweren", GeR 2001: 82, 83) bis zu allen fünf erstreckt (z.B. "Kann in einem Interviewgespräch einfache, direkte Fragen zur Person beantworten, wenn die Fragen langsam, deutlich und in direkter, nichtidiomatischer Sprache gestellt werden", GeR 2001: 85). Eine Systematik ist nicht zu erkennen, jedoch sind Tendenzen auszumachen, und zwar eine durchgehende

Besetzung der Kategorie 'Operation' und eine weniger starke Besetzung der Kategorie 'Einschränkungen' auf den Niveaus C1 und C2.

Eine weitere Art von Inkonsistenz ist in den Beispielen zu beobachten. Beispiele wie "(…) kulturelle Themen, z.B. Musik oder Filme" (GeR 2001: 81) werden in den Deskriptoren nur vereinzelt angegeben. Gerade dies wäre aber nutzerfreundlich und würde die Deskriptoren besser handhabbar machen. Die Beispiele könnten trotzdem allgemein genug bleiben, um für viele Kontexte passend zu sein. In der Anpassung der Deskriptoren mit Hilfe von konkreten Beispielen sehen wir ein großes Potenzial für eine mögliche Adaptierung von Deskriptoren, ohne sich zu weit von den eigentlichen Deskriptoren zu entfernen. Hier ist z.B. das neu erschienene "Europäische Portfolio der Sprachen" (2008) im Grundportfolio einen Schritt in die richtige Richtung gegangen, indem es unterhalb der Ebene der Deskriptoren (weitgehend aus dem GeR) "Indikatoren" formuliert hat, die für die nötige inhaltliche Anbindung sorgen, ohne die Grundschulkinder die jeweiligen Deskriptoren gar nicht verstehen könnten. Auch in "PROFILE Deutsch" (Glaboniat et al. 2005) ist ein Versuch in Richtung einer Konkretisierung unternommen worden, die wiederum andere Probleme aufwirft, die hier nicht erörtert werden können.

Fazit

Als Fazit kann festgehalten werden, dass die Strukturiertheit der Merkmale in der Skala "Mündliche Interaktion allgemein" sowie den zugehörigen Subskalen schwierig zu erkennen ist, weil sie so viele Aspekte involviert. Die Merkmale enthalten eine unsystematische Mischung von Situationen, Funktionen und Diskursformen. Hinsichtlich der Terminologie bestätigt die Analyse der Skalen zur mündlichen Interaktion die Ergebnisse von Alderson et al. (2004) und Harsch (2006, 2007) in anderen Bereichen. Die Verwendung von Synonymen, deren Funktion unklar ist, ist ebenso zu beobachten wie Termini, die unterschiedlich interpretierbar sind. Eine Mischung verbaler und nonverbaler Operationen erfordert Interpretationen, insbesondere auf den niedrigen Niveaus. Eine konsistente, transparente Formulierung von Operationen über Niveaus hinweg fehlt; die Spannbreite von Operationen reicht von konkret zu abstrakt, aber ohne erkennbares System. Begriffe werden nicht definiert und damit unterschiedlich auslegbar.

Die Elemente (*slots*) in den Deskriptoren sind nicht konsistent besetzt. Auch in der Beschreibung des Gegenstands der mündlichen Interaktion ist keine Systematik zu erkennen, wobei pragmatische und funktionale Aspekte in den Skalenbezeichnungen lediglich Anhaltspunkte geben. Man fragt sich, wie sinnvoll es ist, einen solchen "Flickerlteppich" zu skalieren.

Die Autoren des GeR haben diese Problematik durchaus gesehen. In Kapitel 3.7 (GeR 2001: 45) heißt es: "Nicht jedes Element oder jeder Aspekt eines Deskriptors wird auf der jeweils folgenden Ebene wiederholt. Das heißt, dass die Einträge auf jedem Referenzniveau selektiv beschreiben, was als charakteristisch ins Auge fällt." Es werden durchaus plausible Gründe dafür aufgeführt, aber dann heißt es: "Wenn Benutzer des Referenzrahmens die Deskriptoren verwenden wollen, müssen sie sich entscheiden, wie sie mit den Lücken in den angebotenen Deskriptoren umgehen wollen. Es kann durchaus sein, dass man die Lücken füllen kann, indem man für den jeweiligen Verwendungskontext neue Deskriptoren erarbeitet oder vorhandenes Material aus dem System des Benutzers integriert. [...] Eine Lücke in der Mitte einer Skala hingegen kann darauf hinweisen, dass eine sinnvolle Unterscheidung nicht einfach zu formulieren ist" (GeR 2001: 46). Mit diesen salvatorischen Klauseln entlassen die GeR-Autoren aber im Grunde die Nutzer aus der Verantwortung: Wie sollen denn Niveaus vergleichbar bleiben, wenn es jedem freisteht, über die inhaltlichen Füllungen selbst zu befinden? Man kann daraus eigentlich nur noch den Schluss ziehen, dass einzig die Benennungen A1 bis C2 eine Rolle spielen, während die Entscheidungen über die Inhalte arbiträr sind.

5.3 Skalierung und Standards festsetzen

Einige Probleme im Bereich *scaling*, also der Festlegung der Stufen der Referenzniveaus, sind erst dann nachzuvollziehen, wenn man den Prozess der Entstehung der Skalen näher beleuchtet. Die ursprüngliche Referenzskala (es handelte sich um eine Skala, auf der alle Deskriptoren skaliert wurden und die erst später in unterschiedliche Skalen aufgefächert wurde) wurde erstellt auf der Grundlage des sogenannten Raschmodells (vgl. Rasch 1960 / 1980) als einer Variation der *Item Response Theorie*. Es handelt sich hier um eine probabilistische Testtheorie, die Antwortverhalten in Tests (*item responses*) mit formalen Modellen beschreibt (Rost 2004). Das Raschmodell ermöglicht es, Items zu kalibrieren und Personen auf einer gemeinsamen Intervallskala zu messen. Für die Berechnung der gemeinsamen Referenzskala kam ein erweitertes Raschmodell zur Anwendung, das mehrere Facetten beinhalten kann. Die Grundlage dieser Berechnung bieten Antworten auf die entsprechenden Aufgaben, also: Testdaten. Um an diese zu kommen, legte North (2000) Lehrenden u.a. Videoaufzeichnungen von Lernenden vor, die sie mittels Deskriptoren auf Minifragebögen bewerten sollten. Außerdem nutzten sie dieselben Fragebögen für die Bewertung ihrer eigenen Lernenden (n=945). Dieses war der erste Versuch, eine einzige Skala für

potenziell alle Sprachen und für verschiedene Kontexte und Nutzungsarten zu erstellen.

Bei dieser durchaus innovativen Vorgehensweise besteht das Problem darin, dass die Studie die Methodologie zur Erstellung einer Itembank mit einem Datenerhebungsinstrument verbindet, das normalerweise für dichotome Testitems verwendet wird. Daher wird bei letzterem im Normalfall nur ein kleiner Aspekt von Sprachfähigkeit beurteilt und nicht Sprachfähigkeit als Gesamtkonzept. Dieser Umstand könnte gleichzeitig erklären, warum dem Konstrukt keine Theorie von Sprachfähigkeit in ihrer Gesamtheit zugrunde liegt. Ein weiteres methodologisches Problem bestand darin, dass der Analyseprozess inhärente Widersprüche hatte. Itembanking erfordert eine separate Analyse der Daten, während für den anderen Teil (eine FACETS-Analyse) eine integrierte Analyse vonnöten ist. Es ergibt sich damit ein Widerspruch in der Datenerhebungsmethode, der im Verlauf der Untersuchung zu weiteren Problemen bei der Datenanalyse führte, auf die wir hier nicht im Detail eingehen können (vgl. aber Vogt, erscheint 2010).

Diese Widersprüche führen zu Verzerrungen in den Daten, die zwar durch die Adaptierung der Datenanalyseverfahren reduziert werden konnten. Eine vollständige Behebung war nicht möglich, und dies dürfte die Ergebnisse der Datenanalyse entsprechend beeinflusst haben.

An mehreren Stellen mussten subjektive Entscheidungen getroffen werden, die die Qualität der Deskriptoren beeinflussten. Beispielsweise ergaben die Deskriptoren für das Merkmal "Unabhängigkeit" (Carroll 1980), wie oben dargestellt, schlechte Werte, so dass sie eigentlich hätten verworfen werden müssen. Stattdessen wurde deren Inhalt als Proviso formuliert und an geeignete Deskriptoren angehängt. Die Deskriptoren, die das Lesen betrafen, mussten hingegen wegen schlechter Werte aus der Hauptuntersuchung herausgenommen werden und wurden als Notlösung einfach separat analysiert. An anderer Stelle wurden Items, die eigentlich hätten aussortiert werden müssen, beibehalten, so dass unter den Deskriptoren Qualitätsunterschiede zu finden sind, eine Tatsache, die in der Dokumentation des Schweizer Projektes auch vermerkt ist (Schneider & North 2000), nicht aber im GeR selbst.

Die *cut-off points* sind als Ergebnis der Raschanalyse gesetzt worden; North (2000) gibt jedoch selbst zu bedenken, dass beim Setzen von *cut-off points* immer ein subjektives Element hereinspielt. Wie *cut-off points* festgelegt wurden, ist dargestellt in North (2000). In der Rezeption des GeR ist die Skala jedoch als absolut und objektiv angesehen worden. Allgemein ist überhaupt eine kritiklose Akzeptanz der durchaus problematischen statistischen Verfahren bei der Erstellung der (einen!) Allgemeinskala zu verzeichnen. Dies gilt unserer Meinung nach leider auch für das Positionspapier der DGFF.

North (2000: 290f.) macht im Hinblick auf die Progression von Sprachfähigkeit und die Möglichkeiten, diese mit Hilfe von Skalen und Deskriptoren zu dokumentieren, wichtige Einschränkungen, die weder im GeR selbst zu finden sind, noch in der Diskussion und Rezeption des GeR eine Rolle spielen. Er konstatiert, dass es sich bei der Studie um eine objektive (selbst das sei in Anbetracht der vielen Notlösungen im Datenanalyseprozess dahingestellt) Skalierung eines intersubjektiven Konsensus handelt. Keinesfalls erhebt er den Anspruch, dass das Produkt der gemeinsamen Referenzskala und damit das angebotene Bild von Sprachfähigkeit "wahr" ist. Auch gibt er keine Garantie, dass die Skala ein Bild von sich entwickelndem Fremdsprachenerwerb darstellt. Nüchtern betrachtet, erfolgte lediglich eine Passung von Daten (Urteile von Lehrenden über die Sprachfähigkeit Lernender) auf ein statistisches Modell. Aussagen über fremdsprachliche Entwicklungssequenzen kann man damit nicht machen. Es ist daher verfehlt, die Skalen und Deskriptoren im GeR in ihrer Gültigkeit absolut zu setzen (vgl. Vollmer 2003). Genauso wenig dürfen die Skalen und Deskriptoren als eine lineare Messskala interpretiert werden. Der Konsenscharakter der Referenzskala wird von North (2007: 1) explizit bestätigt:

> A 'common framework' like this is a social construct, a constructed consensus. The CEFR descriptors are *shared perceptions* of proficiency – we do not actually know that those perceptions are "correct" and that language proficiency really is as depicted, but they present us with common reference points to discuss things.

Daraus kann gefolgert werden, dass sich der GeR nicht für alle Verwendungszwecke eignet (ausführlich dazu Harsch 2006, 2007), vor allem nicht für die Formulierung von Bildungsstandards. Die Funktion der Skalen und Deskriptoren im GeR als neutrale Referenzfunktion ist nicht mit der Zielsetzung eines normativen Dokuments wie den Bildungsstandards in Einklang zu bringen. Dies erscheint uns ein umso dringlicheres Problem, als die Bildungsstandards an neuralgischen Punkten der Vergabe von Berufs- und Aufstiegschancen inzwischen die dominierende Grundlage der Evaluation darstellen.

5.4 Zu den Defiziten im Bereich der Lerntheorie

Ein großes Problem stellt die fehlende lerntheoretische Basis der Skalen und damit des gesamten Kompetenzbegriffs im GeR dar. Es ist bemerkenswert, mit welcher Nonchalance die Autoren des GeR die letzten Jahrzehnte der Diskussion auf dem Feld der Fremdsprachenlerntheorie beiseite wischen. Nachdem in Kapitel 6.2.1 ("Erwerben oder lernen?") ein kurzer Rückblick auf eine ältere

Auseinandersetzung erfolgt ist, ohne deren Ergebnis korrekt zu benennen, beginnt der Abschnitt 6.2.2.1 mit der Feststellung:

> Es gibt derzeit keinen allgemeinen, auf Forschungsergebnissen basierenden Konsens darüber, wie Lernende lernen; aus diesem Grund kann der Referenzrahmen sich nicht auf eine bestimmte Lerntheorie stützen. ... (GeR 2001: 138)

Dabei wäre eine zeitgemäße psycholinguistische oder allgemein lerntheoretische Basierung dem GeR insgesamt gut bekommen. Stattdessen ziehen sich die Autoren auf ein vages Konzept der "Aufgabenorientierung" zurück, das in Kapitel 7 expliziert wird. Viele damit verbundene Fragen bleiben allerdings offen.

Wie wichtig eine lerntheoretische Basierung dieser Komponente wäre, zeigt ein Blick auf die Deskriptoren. Im Bereich der rezeptiven Fertigkeiten sind die Beschreibungen der Operationen, aus denen "Verstehen" resultiert, nur sehr pauschal erfolgt. Der GeR bezieht sich nicht auf Theorien des Verstehens, wie sie seit Jahrzehnten bekannt sind (etwa das Konzept der "Verarbeitungstiefe" bei Craik & Lockhart 1972, das schon bei Hörmann z.B. 1976 in "Meinen und Verstehen" als Basis einer Typologie der Verstehensakte empfohlen wurde). Alderson et al. (2004) bemängeln dies zu Recht; viele der Ungereimtheiten in den Formulierungen der Deskriptoren wären vielleicht zu vermeiden gewesen, hätte man an diesem wichtigen *slot* des "Moleküls" ein wenig sorgfältiger gearbeitet.

Es wäre eine Herausforderung gewesen, die "kognitiven Operationen", die bestimmten Aktivitäten zugrunde liegen, bei den Paraphrasen des Verbs "verstehen" mit zu berücksichtigen. An Aufgaben, die heute z.B. in reicher Zahl als Indikatoren für das Glücken von Verstehensakten (hier im Bereich Lesen) angesehen werden, kann man zeigen, was in dieser Hinsicht notwendig wäre. Spiegeln z.B. die folgenden Aufgaben auch eine Hierarchie des Verstehens? Welche Verstehensleistungen oder Aktivitäten kann man auf welchen Referenzniveaus erwarten?

- Zwischen den Zeilen lesen, differenzieren, fiktionalen Texten Bedeutung geben
- Texte zusammenfassen (auch übersetzend) / Texte "auf den Punkt bringen"
- Überschriften formulieren / einzelnen Absätzen Überschriften zuordnen
- Textabschnitten passende Zusammenfassungen zuordnen
- nicht passende Texteinschübe identifizieren (und korrigieren)
- Lückentexte ergänzen
- Textabfolgen rekonstruieren (Absätze sortieren)
- Richtig-/Falsch-Entscheidungen treffen (auch Mehrfachwahlantworten)
- Details im Text auffinden und mit parallelen Aussagen vergleichen

- Bilder Texten zuordnen
- Ort / Zeit / Sprecher / Thema identifizieren
- Details identifizieren und markieren.

Diese Aufgaben und andere spiegeln jeweils unterschiedliche kognitive Prozesse, die in den oberen Bereichen der Liste eine größere Verarbeitungstiefe erfordern als in den unteren. Ein Modell der Kompetenzstrukturentwicklung müsste solche Dinge erfassen können; die Niveaubeschreibungen wären bei sorgsamerem Umgang mit den Operatoren (s. o.) sicher aussagekräftiger ausgefallen.

6. Didaktische Implikationen und Ausblick

Wie eingangs beschrieben, hat die KMK in den Bildungsstandards den Schritt gewagt, trotz der sich damals schon abzeichnenden Kritik am GeR Zielniveaus in der Ersten Fremdsprache (Englisch / Französisch) festzulegen, nämlich A2 für die 9. Klasse (in der Regel die Hauptschule) und B1 für die 10. Klasse.

Unsere Diskussion der Probleme bei der Festlegung von *cut-off points* beim *scaling* regt dazu an zu überlegen, ob dies eine weise Entscheidung war. Der Hinweis im GeR (2001: 29), dass die Lernzeit, die man vermutlich benötigt, um von A2 zu B1 zu gelangen, erfahrungsgemäß derjenigen entspricht, die der Schritt von Null bis A2 erfordert, würde bedeuten, dass man Lernenden in dieser Schulform nur "50 Prozent" des Kompetenzniveaus zumutet, das man Realschülern abverlangt, die ja nur ein Jahr länger lernen. Dass die Situation aber noch viel dramatischer ist, zeigt ein Blick in die DESI-Studie (Beck & Klieme 2007; vgl. auch Tschirner 2008). Hauptschüler erreichen offenbar in großer Zahl nur mit Mühe ein Niveau, das A1 im GeR entspricht; die DESI-Studie war allerdings nur indirekt an den Referenzniveaus des GeR orientiert. Solche Beobachtungen werfen die Frage auf, inwieweit nicht eine ganze Reihe personaler und affektiver Faktoren das Erreichen von Referenzniveaus so nachhaltig beeinflussen, dass man die Zielvorstellungen nicht nur nach linguistischen Kriterien formulieren darf.

Die Skalierungen im GeR sind auch in anderer Hinsicht nicht geeignet für das allgemeinbildende Schulwesen in Deutschland. Die inhaltlichen Aspekte der Deskriptoren sind bisweilen so weit weg vom schulischen Bildungsauftrag und so nahe an pragmatischer Kommunikation zwischen Erwachsenen gewählt, dass sich eine Nutzung im Bildungswesen ohne Adaptation und erneute empirische Validierung im Grunde verbietet.

Gerade auf dem Niveau A1 kann man das sehr gut an einem aktuellen "Europäischen Portfolio der Sprachen" studieren, das in einem BLK-Projekt

entwickelt und 2008 von den Schulbuchverlagen Cornelsen, Diesterweg und Klett auf den Markt gebracht wurde. Die Verfasserinnen und Verfasser dieses Portfolios mussten für A1 den Weg gehen, "Indikatoren" meist auf der Ebene von Sprechakten zu formulieren ("Ich kann verstehen, wenn mich jemand nach dem Weg fragt"); wenn eine bestimmte (und überschaubare, im Unterricht vermutlich gelernte) Zahl solcher Indikatoren beherrscht werden (z.B. 8 von 10), kann man sagen, dass der "offizielle" A1-Deskriptor für Hörverstehen eingelöst wird. Das bedeutet, dass vor allem die unteren Niveaus für den schulischen Kontext nicht differenziert genug, vor allem aber nicht zielgruppengerecht formuliert sind. Da in den Begleitdokumenten des GeR insbesondere zu den Europäischen Sprachenportfolios betont wird, dass nur ein empirisch validiertes System eine zuverlässige Basis für neue Deskriptoren abgeben kann (Lenz & Schneider 2004), müsste die Schule entweder ihre Curricula völlig umstellen auf wenig kind- und jugendgerechte pragmatische Inhalte oder aber ein eigenes System von Deskriptoren und Indikatoren schaffen, das wiederum ohne Validierung wenig zuverlässig und für standardisierte Qualifikationen nicht geeignet wäre. Ad hoc formulierte Indikatoren wie im "Europäischen Portfolio der Sprachen" (2008), die sich die Deskriptoren des GeR als Referenzpunkt nehmen, müssten dann kontextspezifische Inhalte und Sprachhandlungen aufgreifen. Sie näherten sich damit aber bereits dem an, was im Bildungswesen als "Lernziel" bekannt ist – und damit könnten auch Lehrkräfte mehr anfangen als mit einem Deskriptorensystem.

Insgesamt muss man feststellen, dass die für Bildungsprozesse im Fremdsprachenunterricht relevanten Aspekte wie die Entwicklung interkultureller Kompetenz (vgl. das Positionspapier der DGFF in der ZFF 2/2008) oder der Bereich des Umgangs mit fiktionalen Texten oder kreativem Sprechen und Schreiben im Grunde so defizitär berücksichtigt werden, dass ein ganz zentraler Aspekt bisheriger Lehrpläne völlig ausgeklammert bleibt. Das Positionspapier der DGFF weist als "Ersatz" für diese Defizite zwar auf das Kompetenzmodell von Weinert (2001) hin – das aber in solchen Fragen wenig hilfreich ist, weil es nicht direkt an die Konkretisierungen der in den Bildungsstandards vorgelegten Kompetenzvorstellungen anzuschließen ist. Diese stammen aus anderen Quellen und Traditionen, und der Bruch zwischen den Zielvorstellungen Weinerts und der DGFF-Autorinnen und -Autoren auf der einen Seite und den pragmatischen Reduktionen auf der anderen ist leider unübersehbar.

Die von GeR und Bildungsstandards ausgelöste Tendenz zu Trivialisierungen sieht man schnell, wenn man die oben zitierten Vergleichsarbeiten für die 6. Klasse mit dem vergleicht, was in älteren wie aktuellen Lehrwerken an Inhalten und Redemitteln bis zur 6. Klasse im Englischunterricht üblich ist, vor allem auch schon im Grundschulenglisch. In dieser Vergleichsarbeit von 2009 findet man wenig altersgemäße Texte wie "Durchsagen am Flughafen verstehen" im Teil

"Hören", vom Blatt gelesene (und überhaupt nicht authentisch gesprochene) Hörtexte wie z.B. den über eine fiktive Schwester, die so beschrieben wird, dass jeder Schulbuchredakteur diesen Text und die dazu gehörende Tonaufnahme abgelehnt hätte. Ein Blick in diese Vergleichsarbeit genügt, um zu sehen, wie die Ausrichtung an GeR und Bildungsstandards zu didaktisch vor allem auch im Sinne eines kommunikativen Fremdsprachenunterrichts inakzeptablen Texten führt. Kritiker wie Bredella (2003) haben sicher recht, wenn sie eine Wende im schulischen Fremdsprachenunterricht in diese Richtung für bedenklich halten.

Den GeR trotz seiner Schwächen zur Basis wichtiger Entscheidungen im Bildungswesen zu nehmen, ist keine sehr glückliche Entscheidung gewesen. Solange er nicht (a) als Dokument des Europarats insgesamt verbessert ist, oder (b) eine für das deutsche Bildungswesen sinnvolle, empirisch validierte Adaptation vorliegt, kann er nur als Referenzobjekt benutzt werden, das Anregungen gibt. Diese Auffassung ist übrigens im GeR immer wieder formuliert: "Die Benutzer sollten prüfen ... / auswählen..." usw. Für die Überarbeitung der Bildungsstandards sollte der GeR als Anregung für eine weitergehende Entwicklung von Kompetenzmodellen und insbesondere prozessorientierten Modellen von Kompetenzentwicklung verwendet werden, mit denen die Ausdifferenzierung von fremdsprachlichen Kompetenzen, die relevant sind für den fremdsprachlichen Bildungsprozess, angegangen werden kann. Sein Stufenmodell ist in der gegenwärtigen Ausprägung nicht für die Festlegung von Kompetenzniveaus in normativen Dokumenten wie den Bildungsstandards geeignet. Eine neue Zuordnung, besser noch eine eigenständig entwickelte Variante mit Blick auf eine möglicherweise völlig veränderte Festschreibung von Niveaus und Qualifikationsprofilen ist bald erforderlich. Vor allem sollte dabei auch der Aspekt berücksichtigt werden, dass Lernende in ihren Kompetenzen "Profile" entwickeln, also z.B. besser sprechen als lesen, oder besser lesen als sprechen können.

Bildungsstandards übernehmen mit der Nutzung der Deskriptoren auch deren Probleme. Es bleibt in der gegenwärtigen Fassung des GeR z.B. zu viel subjektiver Interpretationsspielraum, bedingt durch die fehlende Systematik sowohl in den Skalen als auch innerhalb der Deskriptoren selbst. Der ambige Sprachgebrauch in den Deskriptoren ist für Lehrende wie für Bildungsplaner wenig hilfreich und führt zu unreflektierten Orientierungen. Diese Irritationen sind für ein Bildungssystem eher schädlich.

Wenn das Erreichen der Bildungsstandards durch standardisierte Testsysteme überprüft werden soll, so dürfen diese keineswegs ausschließlich und dogmatisch auf den GeR bezogen sein. Die Arbeit an zentralen Abschlussprüfungen in den Bundesländern, soweit sie uns bekannt ist, zeigt derzeit bedenkliche Tendenzen zu erratischen Entscheidungen. Offenbar werden die gewünschten Standards durch *ad hoc*-Forderungen der Bildungsadministration unterlaufen, die z.B. in zwei uns

bekannten Fällen den Autorinnen und Autoren ihrer Prüfungen auferlegt hat, dass 90 bis 95 Prozent der Schüler "bestehen" müssen – auch wenn ein Zielniveau A2 öffentlich propagiert wird. Dies zeigt, dass die KMK zwar Bildungsstandards haben möchte und den vom GeR hergeleiteten Objektivitäts- und Standardanspruch an sie stellt, in den Ländern selbst aber dies wieder durch wenig standardisierte Abschlussarbeiten konterkariert wird – was sowohl die Bildungsstandards als auch den GeR diskreditiert, weil beide im Grunde nicht ernst genommen werden.

Das IQB geht mit VERA *8* (IQB 2009) in dieser Hinsicht einen plausibleren Weg, indem es für diese Vergleichsarbeiten nicht nur professionell erstellte Aufgaben zu weitgehend authentischen Texten benutzt, sondern auch zumindest ansatzweise offenlegt, wie diese Aufgaben erprobt und standardisiert wurden.

Wenn dabei allerdings nicht nur die Festschreibung der Abschlussniveaus auf A2 bzw. B1 zur Disposition steht, sondern insgesamt auch neue Deskriptoren zur Charakterisierung der geänderten Anforderungen geschaffen werden, wie zu hören ist, käme damit das ganze System der auf dem GeR basierenden regionalen Tests, Lehrmaterialien, Portfolios usw. ins Wanken. Aber vielleicht ist dies ohnehin der einzig gangbare Weg. Der angekündigte *technical report* des IQB zum *standard setting* wird sicher die Probleme der vorschnellen Aneignung des GeR auf den Tisch, und damit weiteren Sprengstoff in die bildungspolitische Diskussion bringen.

Das Grundproblem der Bezüge auf den GeR liegt nämlich darin, dass dessen Autoren zwar immer wieder den reinen Referenzcharakter der Skalen und Deskriptoren betonen, dass diese Skalen aber mittlerweile weltweit eine hohe normative Kraft entfalten. Das geschieht vor allem über Lehrpläne und Bildungsstandards, die das System des GeR dogmatisch benutzen, wiewohl es so nie intendiert war.

Wir fordern daher insbesondere im Kontext von Lehrplan-, Standard- und Testentwicklungen dazu auf, den GeR als Basisdokument kritisch zu rezipieren und seine (teilweise im Dokument selbst formulierten) Grenzen zu berücksichtigen. Eine solche kritische Rezeption stünde gerade einem Verband wie der DGFF gut an, besonders wenn Verfasserinnen und Verfasser des Positionspapiers an der Entstehung der Bildungsstandards bzw. der damit verbundenen Arbeit des IQB selbst beteiligt sind.

Eingang des revidierten Manuskripts: 22.06.2009

Literaturverzeichnis

Alderson, J. Charles (1991), Bands and scores. In: Alderson, J. Charles & North, Brian (Hrsg.), *Language testing in the 1990s: The communicative legacy*. London: British Council / Macmillan, 71-86.

Alderson, J. Charles; Figueras, Neus; Kuijper, Henk; Nold, Günter; Takala, Sauli & Tardieu, Claire (2004), *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: Learning, teaching, assessment. Reading and listening. Final report of the Dutch CEF construct project.* Unveröffentlichtes Manuskript, als Powerpoint. [Online: www.ealta.eu.org/conference/2004/ppt/alderson14may2.ppt. 22.06.2009].

Alderson, J. Charles; Figueras, Neus; Kuijper, Henk; Nold, Günter; Takala, Sauli & Tardieu, Claire (2006), Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly* 3/1, 3-30.

Bachman, Lyle F. (1990), *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bausch, Karl-Richard; Christ, Herbert; Königs, Frank G. & Krumm, Hans-Jürgen (Hrsg.) (2003), *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion.* Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts. Tübingen: Narr.

Beck, Bärbel & Klieme, Eckhard (Hrsg.) (2007), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*, Weinheim: Beltz.

Bredella, Lothar (2003), Lesen und Interpretieren im 'Gemeinsamen europäischen Referenzrahmen für Sprachen': Die Missachtung allgemeiner Erziehungsziele. In: Bausch, Karl-Richard et al. (Hrsg.) (2003), 45-56.

Burwitz-Melzer, Eva & Quetz, Jürgen (2006), Trügerische Sicherheit: Referenzniveaus als Passepartout für den Fremdsprachenunterricht? In: Timm, Johannes-P. (Hrsg.), *Fremdsprachenlernen und Fremdsprachenforschung: Kompetenzen, Standards, Lernformen, Evaluation. Festschrift für Helmut Vollmer*. Tübingen: Narr, 355-372.

Burwitz-Melzer, Eva (2005), Bildungsstandards auf dem Prüfstand (Standards für die Literaturdidaktik). 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, 57-66.

Burwitz-Melzer, Eva (2007), Ein Lesekompetenzmodell für den fremdsprachlichen Literaturunterricht. In: Bredella, Lothar & Hallet, Wolfgang (Hrsg.), *Literaturunterricht, Kompetenzen und Bildung*. Trier: Wissenschaftlicher Verlag Trier, 127-157.

Bygate, Michael (1987), *Speaking.* Oxford: Oxford University Press.

Byram, Michael (1997), *Teaching and assessing intercultural competence*. Clevedon: Multilingual Matters.

Byram, Michael (2008), *From foreign language education to education for international citizenship. Essays and reflections.* Clevedon: Multilingual Matters.

Carroll, Brendan Joseph (1980), *Testing communicative performance: An interim study*. Oxford: Pergamon Press.

Council of Europe (2003), *Relating language examinations to the Common European Framework of Reference: Learning, teaching, assessment (CEF), Manual.* Preliminary pilot version. Strasbourg: Council of Europe.

Council of Europe (2009), *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR), A manual*. Strasbourg: Council of Europe.

Craik, Fergus I.M. & Lockhart, Robert S. (1972), Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, 671-684.

Europäisches Portfolio der Sprachen. Erarbeitet von den Projektgruppen der Länder Berlin, Bremen, Hessen und Nordrhein-Westfalen (2008). Berlin et al.: Cornelsen / Diesterweg / Klett.

Fulcher, Glenn (2004a), Deluded by Artifices? *Language Assessment Quarterly* 1/4, 253-266.

Fulcher, Glenn (2004b), Are Europe's tests being built on an 'unsafe' framework? *The Guardian* (Education), 03.18.2004.

Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen (2001). Hg. Europarat, dt. Übersetzung Quetz, Jürgen et al., erschienen bei München: Langenscheidt und www.goethe.de/referenzrahmen. Engl: Council of Europe (ed.): A Common European Framework of Reference für Languages: Learning, Teaching, Assessment. Cambridge University Press und www.coe.int, hier nur zitiert in der dt. Fassung als GeR 2001).

Glaboniat, Manuela; Müller, Martin; Rusch, Paul; Schmitz, Helen & Wertenschlag, Lukas (2005), *Profile Deutsch. Gemeinsamer europäischer Referenzrahmen. Lernziel-bestimmungen, Kannbeschreibungen, Kommunikative Mittel, Niveau A1-A2-B1-B2-C1-C2*. 2. Aufl. München: Langenscheidt.

Harsch, Claudia (2006), *Der Gemeinsame europäische Referenzrahmen: Leistung und Grenzen. Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt*. Inaugural-Dissertation [Online: http://www.opus-bayern.de/uni-augsburg/volltexte/2006/368/. 22.6.2009].

Harsch, Claudia (2007), *Der gemeinsame europäische Referenzrahmen für Sprachen: Leistung und Grenzen*. Saarbrücken: VDM Verlag Dr. Müller.

Hörmann, Hans (1976), *Meinen und Verstehen*. Frankfurt am Main: Suhrkamp.

House, Juliane (2003), Der Gemeinsame europäische Referenzrahmen für Sprachen – Anspruch und Realität. In: Bausch, Karl-Richard et al. (Hrsg.) (2003), 95-104.

IQ (Institut für Qualitätsentwicklung) Hessen, Hessisches Kultusministerium (2008/2009), *Lernstandserhebung 6*. Aufgabenheft Englisch, Heft B – Listening and Reading und Durchführungsmanual. Unveröffentlichtes Manuskript und CD.

IQB (2009), *VERA 8* – Vergleichsarbeiten 2009, 8. Jahrgangsstufe (VERA 8), Englisch. (Alle Materialien und Handbücher online unter http://www.iqb.hu-berlin.de).

Kleppin, Karin (2003), Der Gemeinsame europäische Referenzrahmen für Sprachen: Ärgernis oder Fortschritt? In: Bausch, Karl-Richard et al. (Hrsg.) (2003), 105-112.

Klieme, Eckhard; Avenarius, Hermann; Blum, Werner; Döbrich, Peter; Gruber, Hans; Prenzel, Manfred; Reiss, Kristina; Riquarts, Kurt; Rost, Jürgen; Tenorth, Heinz-Elmar & Vollmer, Helmut J. (2003), *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Frankfurt: Deutsches Institut für internationale pädagogische Forschung. [Online: http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf. 27.05.2009].

KMK (2003), KMK (2004): Bildungsstandards für die Erste Fremdsprache (Englisch/ Französisch) für den Mittleren Bildungsabschluss (Beschluss vom 04.12.2003), und: Bildungsstandards für die Erste Fremdsprache (Englisch/Französisch) für den Hauptschulabschluss (Beschluss vom 15.10.2004), Herausgegeben vom Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. München et al.: Luchterhand / Wolters Kluwer Deutschland, beide zitiert als "Bildungsstandards".

Kramsch, Claire (1998), The privilege of the intercultural speaker. In: Byram, Michael & Fleming, Michael (Hrsg.), *Language learning in intercultural perspective*. *Approaches through drama and ethnology*. Clevedon: Multilingual Matters, 16-32.

Krumm, Hans-Jürgen (2003), Der Gemeinsame europäische Referenzrahmen – ein Kuckucksei für den Fremdsprachenunterricht? In: Bausch, Karl-Richard et al. (Hrsg.) (2003), 120-126.

Lenz, Peter & Schneider, Günther. (2004), *Introduction to the bank of descriptors for self-assessment in European Language Portfolios*. [Online: http://www.coe.int/T/DG4/Portfolio/?L=E&M=/documents_intro/Data_bank_descriptors.html. 23.05.2009].

Luoma, Sari (2004), *Assessing speaking*. Cambridge: Cambridge University Press.

North, Brian & Schneider, Günther (1998), Scaling descriptors for language proficiency scales. *Language Testing* 15/2, 217-263.

North, Brian (2000), *The development of a common framework scale of language proficiency*. (Theoretical Studies in Second Language Acquisition 8), New York: Peter Lang.

North, Brian (2007), *Symposium: The CEFR in Europe and beyond: Challenges and experiences. Response by Brian North*. [Online: www.ealta.eu.org. 20.05.2009].

Quetz, Jürgen (2002), Der Gemeinsame europäische Referenzrahmen. In: Wolff, Armin; Lange, Martin (Hrsg.), *Europäisches Jahr der Sprachen: Mehrsprachigkeit in Europa*. Materialien Deutsch als Fremdsprache 65, Regensburg: FaDaF, 369-383.

Quetz, Jürgen. (2003), Der Gemeinsame europäische Referenzrahmen: Ein Schatzkästlein mit Perlen, aber auch mit Kreuzen und Ketten. In: Bausch, Karl-Richard et al. (Hrsg.) (2003), 145-155.

Quetz, Jürgen (2004), Der Gemeinsame Europäische Referenzrahmen als ein Modell kommunikativer Kompetenz. In: Quetz, Jürgen & Solmecke, Gert (Hrsg.), *Brücken schlagen. Dokumentation zum 20. Kongress für Fremdsprachendidaktik der DGFF 2003 in Frankfurt am Main*. Berlin: Pädagogischer Zeitschriftenverlag, 211-223.

Quetz, Jürgen (2005), Bildungsstandards, Sprachstandards. In: Bausch, Karl-Richard; Burwitz-Melzer, Eva; Königs, Frank G. & Krumm, Hans-Jürgen (Hrsg.) (2005), *Bildungsstandards für den Fremdsprachenunterricht auf dem Prüfstand.* Arbeitspapiere der 25. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts. Tübingen: Narr, 210-218.

Quetz, Jürgen (2007), Textrezeption im Referenzrahmen, in den Bildungsstandards, in Abschlussprüfungen und im Unterricht. In: Bausch, Karl-Richard; Burwitz-Melzer, Eva; Königs, Frank G. & Krumm, Hans-Jürgen (Hrsg.), *Textkompetenzen.* Arbeitspapiere der 27. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts. Tübingen: Narr, 150-160.

Rasch, Georg (1960 / 1980), *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Rost, Jürgen (2004), *Lehrbuch Testtheorie – Testkonstruktion*. (2. Aufl.) Bern: Hans Huber.

Schmenk, Barbara (2004), Drama in the margins? The Common European Framework of Reference and its implications for drama pedagogy in the foreign language classroom. *GFL German as a Foreign Language* 1, 7-23.

Schneider, Günther & North, Brian (2000), *Fremdsprachen können – was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit.* Chur: Rüegger.

Tönshoff, Wolfgang (2003), Referenzrahmen: Zwischen Ansprüchen und Erwartungen. In: Bausch, Karl-Richard (Hrsg.) (2003), 180-191.

Tschirner, Erwin (2008), Vernünftige Erwartungen: Referenzrahmen, Kompetenzniveaus, Bildungsstandards. *Zeitschrift für Fremdsprachenforschung* 19/2, 187-208.

Vogt, Karin (2007), Anpassung von Skalen und Deskriptoren des Gemeinsamen Europäischen Referenzrahmens am Beispiel des berufsorientierten Fremdsprachenlernens: das Forschungsprojekt Kompetenzprofile. *Zeitschrift für Fremdsprachenforschung* 18/1, 1-24.

Vogt, Karin (Habilitationsschrift, erscheint 2010), *Fremdsprachliche Kompetenzprofile: Entwicklung und Abgleichung von GER-Deskriptoren für Fremdsprachenlernen mit einer beruflichen Anwendungsorientierung*.

Vollmer, Helmut J. (2003), Ein gemeinsamer europäischer Referenzrahmen für Sprachen: Nicht mehr, nicht weniger. In: Bausch, Karl-Richard et al. (Hrsg.) (2003), 192-206.

Weinert, Franz E. (2001), Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: Weinert, Franz E. (Hrsg.), *Leistungsmessungen in Schulen*. Weinheim: Beltz, 17-31.