Der GeR als Referenzsystem für kompetenzorientiertes Testen: Was bedeutet der Bezug zum GeR für eine Sprachprüfung?

Gabriele Kecker[1]

Since its publication in 2001, the Common European Framework of Reference (CEFR; Council of Europe 2001) has become a widely recognized standard of competence level descriptions for languages. Educational standards, textbooks, language courses, and language examinations in European countries and beyond have been developed with reference to the CEFR scales. This paper describes which expectations of the CEFR as a reference system can be considered as realistic and which quality standards must be fulfilled by language examinations in order to be able to establish a reliable link to the CEFR. With regard to centralized language tests which currently need to be developed for school leaving examinations in many federal *Länder* in Germany, different approaches for validating an alignment to the CEFR and their possible implementations are discussed.

1. Einleitung

Die Evaluation von Leistungen erfordert unabhängig von ihrer Form eine Bezugsgröße, mit deren Hilfe Indikatoren für die betreffende Kompetenz entwickelt werden können. Diese Bezugsgröße wird im Allgemeinen durch Curricula oder Lehrpläne vorgegeben, die eine Beschreibung der zu erreichenden Kompetenzen und der gewünschten Kompetenzstufe enthalten. Als Grundlage des Curriculums oder Lehrplans werden häufig Bildungsstandards oder andere Referenzsysteme herangezogen. Ein mittlerweile sehr verbreitetes Beispiel solcher Referenzsysteme für das Erlernen von Fremdsprachen ist das *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, CEFR (= *Gemeinsamer europäischer Referenzrahmen für Sprachen*, GeR), der im Jahr 2001 vom Europarat veröffentlicht wurde (Europarat 2001). Andere Referenzsysteme zur Beschreibung von Sprachkompetenz wie die *Proficiency Guidelines* des *American Council on the Teaching of Foreign Languages* (ACTFL) wurden bereits 1986 in den USA entwickelt (2012 revidiert) oder im Fall der ALTE-Niveaustufenskalen seit Mitte der neunziger Jahre von dem Verband der Sprachtester in Europa (*Association of Language Testers in Europe,* ALTE) zur Verfügung gestellt.

1 Korrespondenzadresse: Gabriele Kecker, TestDaF-Institut, Universitätsstr. 134, 44799 Bochum, Tel. 0234-3229720, Gabriele.Kecker@testdaf.de

1.1 Funktion von Referenzsystemen

Referenzsysteme wie die zuvor genannten Beispiele erfüllen unterschiedliche Funktionen, die im Folgenden beschrieben werden sollen (einen Überblick zur Verwendung des GeR bieten Alderson 2002; Galaczi & Weir 2013). Sie dienen als Bezugspunkt für die im Unterricht zu vermittelnde Kompetenz und das in der Leistungsmessung zu messende Konstrukt. So ist mit der Verwendung des GeR häufig die Absicht verbunden, aufgrund des handlungsorientierten Ansatzes kommunikativer Kompetenz eine stärkere Kompetenz- und Output-Orientierung im Fremdsprachenunterricht einzuführen. Dementsprechend wurden beispielsweise in Deutschland von der Kultusministerkonferenz (KMK) Bildungsstandards für den Mittleren Schulabschluss und das Abitur (KMK 2003, KMK 2012) in den Fremdsprachen Englisch und Französisch entwickelt, die einen starken Bezug zum GeR aufweisen und die intendierten bildungs- und sprachpolitischen Ziele implementieren sollen. Ähnliche Entwicklungen sind auch in Österreich und der Schweiz (Projekt HarmoS) zu verzeichnen (vgl. BIFIE 2012; Konsortium HarmoS Fremdsprachen, Schneider, Lenz & Studer 2009). Neben diesen eher bildungspolitischen Zielen wird mit der Verwendung von Referenzsystemen noch ein anderer Zweck verfolgt: Sie geben allen beteiligten Interessengruppen (*stakeholders*), d.h. zunächst den Lernenden, ihren Eltern sowie den Lehrkräften, aber z.B. auch Arbeitgebern, Zulassungsstellen oder anderen Entscheidungsträgern ein möglichst genaues Bild der Kompetenz, die es zu erwerben und zu vermitteln gilt. Wird diese Kompetenz mit einer Note bewertet, so ist häufig unklar, wie die damit verbundene Leistung aussieht, d.h. wie die Note zu interpretieren ist. "We can add meaning to the scores by referencing them to […] performance levels, benchmark performance levels, or achievement levels (e.g., as in [...] CEFR)" (Kane 2012: 8). Im Unterschied zu rein numerischen Bewertungen haben Kompetenzbeschreibungen, insbesondere in der Form von Can-do-Statements, den Vorteil, dass sie aussagekräftiger sind und von den Beteiligten leichter verstanden werden. Dies setzt jedoch eine möglichst eindeutig formulierte Beschreibung der Kompetenz voraus und zusätzlich eine einheitliche Interpretation durch den jeweiligen Nutzer. Beide Aspekte, d.h. eine eindeutige Kompetenzbeschreibung und eine konsensuelle und standardisierte Interpretation bilden die Voraussetzung für eine konsistente Verwendung über europäische Grenzen hinweg. Diese gewinnt an Bedeutung, wenn Schulabschlussnoten oder Ergebnisse von Sprachprüfungen,[2]

2 Die Begriffe "Test" und "Prüfung" werden im vorliegenden Beitrag synonym verwendet. "Unter Test oder Prüfung soll jegliches Verfahren gefasst werden, das Individuen unter kontrollierten Bedingungen zu bestimmten Handlungs- und Verhaltensweisen veranlasst, die wiederum Rückschlüsse ermöglichen sollen auf bestimmte Fähigkeiten und Fertigkeiten" (Grotjahn 2003: 9; vgl. auch Davies, Brown, Elder, Hill, Lumley & McNamara 1999: 56-57).

die durch Kompetenzbeschreibungen verdeutlicht werden sollen, weitreichende Konsequenzen haben. Vor dem Hintergrund der Erwartungen an die Kompetenz der betroffenen Personen und der Konsequenzen, die z.B. mit einer Abiturnote oder dem Ergebnis einer Sprachprüfung verbunden sind, wird zusätzlich deutlich, dass der Bezug zum Referenzsystem sowie die Prüfung selbst hohen Qualitätsanforderungen entsprechen müssen. Anderenfalls wären die Entscheidungen von Arbeitgebern, Zulassungsstellen für Studienplätze oder Ausländerbehörden auf dieser Grundlage nicht zu rechtfertigen. Die Anbieter von Sprachprüfungen, seien es privatrechtliche Einrichtungen oder auch Bildungsbehörden, sind daher gehalten, die Zuordnung ihrer Prüfungen zu einem Referenzsystem wie dem GeR in geeigneter Form nachzuweisen, d.h. die Interpretation des Prüfungsergebnisses mit GeR-Bezug zu validieren.

1.2 Zum Begriff Validität

Sofern Prüfungsanbieter eine Validierung des GeR-Bezugs ihrer Prüfungen anstreben, ist es erforderlich sich mit dem Begriff Validität auseinander zu setzen, um die notwendigen Nachweise für eine Validierung identifizieren zu können. Im Folgenden soll daher das in diesem Beitrag zugrunde gelegte Verständnis von Validität kurz erläutert werden.

Validität wird in der einschlägigen Literatur seit Jahren nicht mehr als Merkmal angesehen, das mit einer Prüfung oder einem Test selbst in Verbindung gebracht wird (eine Sprachprüfung ist valide), sondern der Begriff Validität wird auf das Ergebnis einer Sprachprüfung bezogen, das im Hinblick auf den zuvor definierten Verwendungskontext interpretiert wird (z.B. ein erfolgreicher Absolvent einer B1-Prüfung kann in der Realsituation entsprechend der Konstruktbeschreibung bestimmte sprachliche Handlungen vollziehen). Im Gegensatz zu früheren Jahren wird dabei nicht mehr zwischen unterschiedlichen Arten der Validität unterschieden (Inhaltsvalidität, Kriteriumsvalidität, Konstruktvalidität), sondern seit Messick (1989) wird Konstruktvalidität als übergeordnetes einheitliches Konzept angesehen, unter dem alle anderen Arten von Validität subsumiert werden (vgl. Kecker 2011: 22f.).

> Construct validity also subsumes content relevance and representativeness as well as criterion-relatedness, because such information about the content domain of reference and about specific criterion behaviors predicted by the test scores clearly contributes to score interpretation (Messick 1989: 17).

Des Weiteren fungiert Validität nicht mehr als absolute Eigenschaft, sondern als Kontinuum, das je nach Aussagekraft der Nachweise und je nach Datenlage mehr

oder minder zutrifft. Dementsprechend wird die Validierung einer Ergebnisinterpretation und der damit verbundenen Inferenzen zu einem kontinuierlichen Prozess, der durch unterschiedliche Untersuchungen gestützt und vorangetrieben wird. Die hier aufgeführten Merkmale des Begriffs Validität bilden nach Eckes (2015b: 451) "den konsensuellen Kern des gegenwärtig vorherrschenden Verständnisses von Validität". Andere Schwerpunkte in der Definition des Begriffs und deren Diskussion in Expertenkreisen können hier nicht erörtert werden (vgl. dazu z.B. Eckes 2015b; Kane 2006).

1.3 Validität und der Bezug zu dem GeR

Übertragen auf die Zuordnung von Sprachprüfungen zu einem Referenzsystem wie dem GeR folgt daraus, dass Inferenzen, die von Prüfungsergebnissen abgeleitet werden (hier die Sprachkompetenz auf einer bestimmten GeR-Stufe), durch möglichst viele unterschiedliche empirische Nachweise abgesichert sein sollten, damit die Inferenzen als valide bezeichnet werden können. Dies gilt insbesondere für *high-stakes tests* und schließt andere mögliche Bestandteile des Konstrukts außerhalb des GeR-Bezugs mit ein. Die Durchführung einer solchen Validierung erfordert neben finanziellen und personellen Ressourcen auch entsprechende Expertise, die nicht in allen Bildungseinrichtungen zur Verfügung steht. Der Europarat hat bereits wenige Jahre nach Veröffentlichung des GeR den Bedarf an Know-how in diesem speziellen Bereich erkannt und durch die Veröffentlichung des *Manuals: Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe, kurz CoE 2009) berücksichtigt. In diesem *Manual* wird von den Autoren ein methodischer Ansatz in fünf Schritten dargelegt, mit dem die Zuordnung einer Sprachprüfung zum GeR auf eine empirische Grundlage gestellt werden kann. Bereits in der Einleitung des *Manuals* weisen die Autoren darauf hin, dass Testanbieter möglichst alle methodischen Schritte des Verfahrens umsetzen sollten:

> A claim that a qualification is linked to the CEFR can only be taken seriously if evidence exists that claims based on specifications (content standards) and standard setting (performance standards) are corroborated through validation (CoE 2009: 13-14).

Familiarisation und *standardisation/benchmarking* sind im Zitat nicht aufgeführt, sie gelten als vorbereitende Phasen für *specification*, *standard setting* und *empirical validation.* Dennoch scheint es für viele Institutionen im Schulsektor schwierig, diesen Qualitätsanspruch einzulösen. Der im *Manual* beschriebene methodische Ansatz entspricht dem bereits dargestellten Verständnis von Validität,

entzieht sich jedoch einem schnellen Zugriff und verlangt Zeit sowie eine tiefergehende Beschäftigung mit der vorgestellten Methode. Eine weitere Hürde bei der Umsetzung des Verfahrens zur Anbindung an den GeR ergibt sich durch die Qualitätsanforderungen an die Sprachprüfung selbst, sofern diese nicht im Großen und Ganzen erfüllt werden. Vor diesem Hintergrund soll in diesem Beitrag untersucht werden, welche Schritte des methodischen Ansatzes aus dem *Manual* für die GeR-Zuordnung von Sprachprüfungen umsetzbar sind (z.B. für Schulabschlussprüfungen) oder welche alternativen, beispielsweise statistischen Verfahren, ebenfalls zu zuverlässigen Ergebnissen der Anbindung an den GeR führen würden.

Aus dem zuvor Gesagten ergibt sich eine Reihe von Fragen, die die Problematik der Verwendung des GeR als Referenzsystem für die Messung und Bewertung von Sprachkompetenz offen legen:

1. Welche Erwartungen an den GeR als Bezugsrahmen für Sprachkompetenz sind realistisch? Sind die GeR-Skalen eindeutig formuliert und bieten sie eine ausreichend präzise Beschreibung der kommunikativen Aktivitäten, um eine einheitliche Interpretation durch Nutzer zu gewährleisten?
2. Welche Anforderungen/Voraussetzungen müssen Sprachprüfungen erfüllen, um den Bezug zu einer GeR-Stufe zuverlässig nachzuweisen zu können?
3. Welche Verfahren der Validierung sind z.B. für Bildungsbehörden in deutschen Bundesländern anwendbar, um den Bezug ihrer Sprachprüfungen zum GeR zu stützen?

Im Beitrag wird zunächst dargestellt, inwieweit sich der GeR als Bezugsgröße für die Entwicklung und Validierung von Sprachprüfungen eignet. Ergänzend dazu werden die Qualitätsanforderungen erläutert, die Sprachprüfungen erfüllen sollten, damit sie einem Referenzsystem zugeordnet werden können. Zum Abschluss werden Methoden erörtert, die zur Validierung eines GeR-Bezugs geeignet sind und unter bestimmten Voraussetzungen auch von Bildungsinstitutionen aus dem Schulsektor umgesetzt werden könnten. Dabei wird auch der methodische Ansatz im *Manual* des Europarats berücksichtigt.

2. Der GeR als Bezugssystem für die Entwicklung und Validierung von Sprachprüfungen

Der GeR wurde in den Jahren 1993 bis 1996 im Auftrag des Europarats mit der Intention entwickelt, "eine gemeinsame Basis [...] für die Entwicklung von zielsprachlichen Lehrplänen, curricularen Richtlinien, Prüfungen, Lehrwerken usw. in ganz Europa" (Europarat 2001: 14) zu schaffen. In dieser Hinsicht wurde der GeR seit seinem Erscheinen zum wichtigsten Bezugspunkt für die Arbeit von europäischen Bildungsbehörden, Lehrkräften, Testentwicklern und Lehrwerksautoren. Neben den unbestrittenen Vorteilen des GeR, zu denen beispielsweise die empirische Kalibrierung der meisten GeR-Skalen in einem Schweizer Projekt gehört (vgl. Harsch 2014: 156; Schneider & North 2000), wurden auch Schwachpunkte identifiziert, die Kritik hervorgerufen haben und die hier nur kurz umrissen werden können. Darunter fällt die ausschließliche Ausrichtung des GeR auf erwachsene Lernende, die fehlende Berücksichtigung von Erkenntnissen der Spracherwerbstheorie und Sprachlerntheorie im Konzept (Harsch 2005; Kleppin 2003), die fehlende Validierung der Skalen mithilfe von Lernerkorpora (Hulstijn 2007), der Widerspruch zwischen den intendierten kontextneutralen und kulturunabhängigen Kompetenzbeschreibungen in den Niveaustufenskalen einerseits und ihrer ausschließlichen empirischen Validierung mit Schweizer Lehrkräften und keinen anderen europäischen Probanden andererseits (Jones 2013: 107). Insgesamt bieten die Skalen laut Alderson (2007: 23) ein Lernerportrait oder eine Momentaufnahme, die der Wahrnehmung der am Schweizer Projekt beteiligten Lehrkräfte von Sprachkompetenz entspricht.

Der Beschreibung der Kompetenzniveaus in den GeR-Skalen und ihrer Verwendung für die Testentwicklung wurde von den Autoren des GeR eine besondere Bedeutung zugewiesen:

> Eines der Ziele des Referenzrahmens ist es, allen beteiligten Partnern bei der Beschreibung der Kompetenzniveaus zu helfen, die gemäß den Standards ihrer Tests und Prüfungen erwartet werden. Dies soll den Vergleich zwischen verschiedenen Qualifikationssystemen erleichtern. Zu diesem Zweck sind ein Beschreibungssystem und die Gemeinsamen Referenzniveaus entwickelt worden (Europarat 2001: 32).

Der zitierte Abschnitt geht davon aus, dass der gemeinsame Bezug von Sprachprüfungen oder Tests auf den GeR eine größere Vergleichbarkeit erlaubt. Sowohl das deskriptive System als auch die Niveaustufenskalen des GeR bieten Anhaltspunkte für eine globale Beschreibung der Inhalte und Kompetenzen, reichen jedoch nicht aus, um als Grundlage für eine sorgfältige Testentwicklung und Testanalyse zu dienen. Angaben zu Lebensbereichen, Themen und kommunikativen Aufgaben sowie Aktivitäten (Europarat 2001: Kap. 4) können als Ausgangs-

punkte in der Testentwicklung verwendet werden, um eine eigene kontextbezogene inhaltliche Beschreibung des Tests zu erstellen. Die Zielsetzung und das Format von Testaufgaben und Test-Items können hingegen auf dieser Grundlage nicht hinlänglich erfasst werden. Hinsichtlich der GeR-Skalen sieht Alderson (2007: 26) den Grund dafür darin, dass sie vorwiegend als nutzerorientierte Skalen konzipiert worden sind und nicht als Skalen für Testentwickler.

Die Eignung der GeR-Skalen für die Testentwicklung wurden beispielsweise von der *Dutch Construct Group* (Alderson, Figueras, Kuijper, Nold, Takala & Tardieu 2006) sowie von Harsch (2005), Kecker (2011) und auch Wisniewski (2014) analysiert und ausführlich dargestellt. Ich möchte mich in diesem Beitrag auf folgende Punkte beschränken, die für die Zuordnung einer Sprachprüfung oder eines Tests zum GeR eine grundlegende Rolle spielen: a) Abstraktheit der Kompetenzbeschreibungen, b) fehlende Parameter für die Testentwicklung, c) mangelnde Kohärenz der Skalen.

a) Die Skalen zu den kommunikativen Aktivitäten des GeR (Europarat 2001: Kap. 4) bieten wenig Hinweise zu mentalen Prozessen der Sprachverarbeitung, die zum Testkonstrukt gehören und für die Testentwicklung genauer beschrieben werden müssen, z.B. die Verarbeitungstiefe beim Verstehen von schriftlichen oder mündlichen Texten (Alderson et al. 2006; Weir 2005). Projekte wie DIALANG, ein diagnostischer Online-Test zur Selbsteinschätzung mit GeR-Bezug, haben zu der Erkenntnis beigetragen, dass die GeR-Skalen insgesamt eine relativ abstrakte Beschreibung sprachlicher Leistungen bieten, die von Testentwicklern mit konkreten Inhalten zu füllen sind (Alderson 2005, 2007). Dazu gehören Bezeichnungen in den Deskriptoren der Skalen wie beispielsweise "einfach", "vertraut", "komplex" oder "kann sich verständigen", "kann sich ausdrücken" in Kombination mit Texten oder Äußerungen. Bei der Verwendung und Interpretation solcher Angaben durch unterschiedliche Nutzer entstehen häufig heterogene Ergebnisse. Insofern kann das zuvor angeführte Ziel der Vergleichbarkeit von Sprachprüfungen durch den Bezug auf gemeinsame Kompetenzbeschreibungen nur schwer erreicht werden.

b) Des Weiteren fehlen wichtige Kontextparameter für die Testentwicklung wie Handlungsabsicht, zeitliche Begrenzungen oder das Antwortformat, welche die intendierten Sprachhandlungen und die damit verbundenen Verarbeitungsprozesse vereinfachen oder erschweren können. Lediglich Sprechgeschwindigkeit und Artikulation gesprochener Sprache werden genannt. Auch der visuelle Input für Testaufgaben wie Diagramme, Ablaufschemata oder Fotos kann in den Skalen nicht verortet werden. Das Betrachten von Filmen oder Videos als solches wird jedoch in einer gesonderten Skala behandelt.

c) Manche Skalen (z.B. "Leseverstehen allgemein", Europarat 2001: 74) beinhalten auf einer Niveaustufe lediglich Deskriptoren mit qualitativen Angaben, die zwar im Unterricht die Einstufung der Kompetenz unterstützen, in Testverfahren jedoch nicht sinnvoll sind, weil man sie kaum beobachten kann (B2: "Kann sehr selbstständig lesen, Lesestil und -tempo verschiedenen Texten und Zwecken anpassen …"). Darüber hinaus ist in vielen Skalen die obere Stufe C2 oder sogar C1 nicht belegt: beispielsweise fehlen C1 und C2 in den Skalen "Zusammenhängendes monologisches Sprechen: Argumentieren" (ibd.: 65) und "Informationsaustausch" (ibd.: 83). Häufig wird nur auf die darunter liegende Stufe verwiesen.

Die zuvor genannten Punkte erschweren es, Testaufgaben und die dazugehörige Leistung mithilfe der GeR-Skalen zu beschreiben und zur Niveaueinstufung auf den Skalen zu verankern, insbesondere, wenn es sich um Material für Sprachprüfungen oder Tests auf den GeR-Stufen C1 oder C2 handelt.

Vor dem Hintergrund dieser Kritik und dem wachsenden Bedürfnis von Testanbietern, einen Bezug ihrer Sprachtests zum GeR in geeigneter, nachvollziehbarer Form nachzuweisen, wurden daher vom Europarat verschiedene Instrumente entwickelt und Projekte initiiert, um einerseits die Kompetenzbeschreibungen auf den GeR-Niveaustufen zu illustrieren und andererseits Testanbieter bei der GeR-Zuordnung zu unterstützen:

– Detaillierte Kriterienraster (CEFR *Content Analysis Grid for Listening & Reading*, CEFR *Content Analysis Grid for Speaking*, CEFR *Content Analysis Grid for Writing*) zur Analyse und Beschreibung von Testaufgaben in den vier verschiedenen Teilkompetenzen, die für die rezeptiven Teilkompetenzen im Rahmen des *Dutch CEF Construct Projects*[3] entwickelt wurden (Alderson et al. 2004, 2006). Die Raster für die produktiven Teilkompetenzen wurden von den Mitgliedern der ALTE *Manual Special Interest Group* entwickelt und dem Europarat zur Verfügung gestellt. Im Folgenden werden die Kriterienraster als CEFR-Grids bezeichnet.

– Eine vom Europarat herausgegebene CD-ROM (CoE 2005) mit Beispielen für Testaufgaben im Lese- und Hörverstehen in fünf europäischen Sprachen (Deutsch, Englisch, Französisch, Italienisch, Spanisch) auf den sechs Stufen des GeR, die mit den Kriterien des CEFR-Grid beschrieben worden sind. Diese CD wird derzeit mit neuen Beispielen ergänzt und soll 2016 neu erscheinen.

3 Ein Projekt, das vom niederländischen Ministerium für Bildung, Kultur und Wissenschaft mit dem Ziel finanziert wurde, ein GeR-basiertes Instrument zu entwickeln, das Konstrukte von Testaufgaben und Tests im Lese- und Hörverstehen beschreibt (vgl. Alderson, Figueras, Kuijper, Nold, Takala & Tardieu 2004: 1).

- Eine DVD mit Beispielen mündlicher Produktion und Interaktion von jugendlichen Lernenden im Alter von 13-18 Jahren in den o. g. fünf Sprachen. Diese Leistungsbeispiele wurden im Rahmen einer internationalen Konferenz in Sèvres im Juni 2008 mithilfe von 48 Experten bewertet und den sechs Niveaustufen des GeR zugeordnet. Ein Bericht zum verwendeten Verfahren und zur Auswertung der Expertenbewertungen ist auf der Webseite des Europarats erhältlich (Breton, Lepage & North 2008).
- Das Projekt *Relating Language Examinations to the Common European Framework of Reference for Languages,* das von 2004 bis 2008 vom Europarat mit dem Ziel durchgeführt wurde, ein Handbuch (das sog. *Manual*, CoE 2009) mit einem methodischen Ansatz zu entwickeln, der dazu dient, Sprachprüfungen valide dem GER zuzuordnen.

Die zuvor beschriebenen Bestrebungen des Europarats, den GeR mit zusätzlichen Instrumenten für Testentwicklung und -analyse zu vervollständigen, verfolgen nicht das Ziel, eine sachgerechte und professionelle Vorgehensweise bei der Entwicklung einer Sprachprüfung oder eines Tests zu ersetzen. Diese sieht entsprechend der einschlägigen Referenzliteratur (vgl. Bachman & Palmer 2010; Downing & Haladyna 2006; Fulcher & Davidson 2007) vor, dass ein Testkonstrukt in erster Linie auf den spezifischen Verwendungskontext, die Zielsetzung und die Zielgruppe zugeschnitten sein muss. Erst in einem zweiten Schritt sollte überprüft werden, inwieweit das Testkonstrukt und die Inferenzen, die aus den Testergebnissen abgeleitet werden können, mit bestimmten Kompetenzstufen des GeR übereinstimmen. Leider wurde gerade in den ersten Jahren nach Veröffentlichung des GeR die Testentwicklung eher normorientiert von den kontextneutralen GeR-Kompetenzbeschreibungen in den Skalen abhängig gemacht. Das heißt, kontextspezifische Sprachhandlungen und Settings wurden in Tests kaum abgebildet, wenn sie sich nicht von vornherein dem GeR zuordnen ließen. Jones und Saville (2009) sehen diese Entwicklung als zu restriktiv an: "[...] some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring a context to the CEFR" (Jones & Saville 2009, zitiert nach Milanovic 2009: 3). Davon abgesehen entspricht ein solches Vorgehen nicht der ursprünglichen Intention, mit der der GeR veröffentlicht wurde. In dem Zusammenhang sollte ebenfalls nicht vergessen werden, dass die Kompetenzskalen keine Definition von Sprachkompetenzen bieten, sondern von den Autoren des GeR als Illustration, das heißt als Beispiel gedacht sind, die in einer früheren Entwurfsfassung des GeR sogar im Anhang aufgeführt war.

Sofern Kontextspezifika bei einer Testentwicklung berücksichtigt werden, dürfen daher keine allzu hohen Erwartungen an die Vergleichbarkeit von Sprachprüfungen mit GeR-Bezug gestellt werden. Zwei Sprachprüfungen, die beispielsweise beide

auf der GeR-Nivaustufe B1 angesiedelt sind, werden beide entsprechende Merkmale für diese Stufe aufweisen, müssen jedoch nicht zwangsläufig das gleiche Konstrukt beinhalten und dieselben Kompetenzen abprüfen (vgl. Harsch 2015: 497). Auch die Zuordnung zum GeR müssen wir uns aus den zuvor genannten Gründen eher als partielle Annäherung und nicht als genaue und vollkommene Übereinstimmung vorstellen.

3. Qualitätsanforderungen an Sprachprüfungen mit Bezug zu einem Referenzsystem

In Abschnitt 1 wurde bereits dargelegt, welche Erwartungen an die Interpretation von Testergebnissen geknüpft werden, die einen GeR-Bezug beinhalten und welche Qualitätsansprüche damit verbunden sind, sofern es sich um *high-stakes tests* handelt. Die Autoren des *Manuals* (CoE 2009) fassten ihre Qualitätsanforderungen an eine solche Sprachprüfung folgendermaßen zusammen:

> Linking a test to the CEFR cannot be valid unless the examination or test that is the subject of the linking can demonstrate validity in its own right. A test that is not appropriate to context will not be made more appropriate by linking to the CEFR; an examination that has no procedures for ensuring that standards applied by interviewers or markers are equivalent in severity, or that successive forms of tests administered in different sessions are equivalent, cannot make credible claims of any linkage of its standard(s) to the CEFR because it cannot demonstrate internal consistency in the operationalization of its standard(s) (ibd.: 9).

Demzufolge erachteten es auch die Autoren des *Manuals* für sinnvoll, in einem ersten Schritt eine dem jeweiligen Kontext angemessene Prüfung zu entwickeln, die dem Bedarf der intendierten Zielgruppe und dem Verwendungszweck entspricht und erst in einem zweiten Schritt zu prüfen, inwieweit sich diese Prüfung einem Referenzsystem wie dem GeR zuordnen lässt (vgl. Abschnitt 2). Dies bedeutet, zunächst das Konstrukt und den Inhalt einer Sprachprüfung oder eines Sprachtests zu beschreiben, das Testformat und die Testauswertung zu entwickeln sowie die Untersuchungsmethoden auszuwählen, mit denen die Qualität der Items und Aufgaben empirisch nachgewiesen werden kann. Weiterhin gehören dazu Untersuchungen zur Stabilität und Konsistenz der Testergebnisse über verschiedene Testversionen hinweg und zur Validität des Konstrukts. In der Fachliteratur finden Testanbieter Beschreibungen der einschlägigen Methoden und Instrumente, die für einen solchen Qualitätsnachweis benötigt werden (vgl. beispielsweise Bachman & Palmer 2010; Downing & Haladyna 2006; Fulcher & Davidson 2007). Auch in den Richtlinien und Qualitätsstandards von europäischen und internationalen Testorganisationen sind entsprechende Darstellungen von *good practice* niedergelegt

(vgl. ALTE – *Association of Language Testers in Europe*, www.alte.org; EALTA – *European Association of Language Testing and Assessment*, www.ealta.eu.org; ILTA – *International Language Testing Association*, www.ilta.org, ITC – *International Test Commission*, www.intestcom.org).

Qualitätsnachweise zur Testentwicklung wie die zuvor dargestellten werden vorwiegend von Anbietern großer standardisierter *high-stakes tests* oder im Rahmen von Schulleistungsstudien durchgeführt, in weitaus geringerem Maße jedoch von Schulen oder Schulbehörden, die die Verantwortung für Schulabschlussprüfungen wie dem Abitur innehaben. Zwar wurde der Kompetenzentwicklung von Lehrkräften im Bereich *assessment literacy* in den vergangenen zehn Jahren in Deutschland mehr Aufmerksamkeit gewidmet, beispielsweise durch die Beteiligung an der Entwicklung von Lern- und Testaufgaben des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) zur Implementierung der nationalen Bildungsstandards der Kultusministerkonferenz (vgl. Porsch, Tesch & Köller 2010a). Dennoch wurde keine systematische Umsetzung von internationalen Qualitätsstandards in die schulische Leistungsmessung eingeführt, etwa durch eine gezielte Aus- und Weiterbildung der Lehrkräfte oder durch Testentwicklung für Schulabschlussprüfungen mittels staatlicher Organisationen wie dem IQB in Deutschland oder Cito in den Niederlanden. Cito übernimmt in den Niederlanden die Testentwicklung für die Schulabschlussprüfungen aller Schulstufen in den Fremdsprachen entsprechend den Qualitätsanforderungen internationaler Standards.

Im Schulsektor in Deutschland wird für fehlende Qualitätsstandards häufig als Begründung angegeben, dass Testaufgaben für Schulabschlussprüfungen (es geht hier nicht um informelle Tests für formative oder summative Leistungsmessung im Unterricht) etwa vor dem Einsatz nicht erprobt werden können, um die Testsicherheit nicht zu gefährden. Dennoch können Qualitätsstandards wie die Testgütekriterien Validität, Reliabilität und Objektivität in gewissem Umfang auch bei der Entwicklung von Sprachprüfungen für den Schulabschluss eingehalten werden, wie im Folgenden dargestellt werden soll.

a) Testspezifikation

Der erste Schritt in der Testentwicklung besteht in der Festlegung des Konstrukts und der theoretischen Definition der Kompetenzbereiche, die gemessen werden sollen. Darauf folgt eine detaillierte Beschreibung der zum Kontext passenden und zu prüfenden Inhalte und Aufgaben, die dem Sprachmodell entsprechen, das dem Test zugrunde liegt. Häufig wird dieses von Bildungsstandards oder einem Curriculum vorgegeben. Die genaue Beschreibung dient der Standardisierung, denn alle mit der Aufgabenentwicklung befassten Personen sollten dieselbe Testspezifikation verwenden, um eine einheitliche Grundlage für die Aufgabenentwicklung zu gewährleisten. Des Weiteren kann

zu einem späteren Zeitpunkt auf der Grundlage der Testspezifikation die Validierung des Testkonstrukts anhand der daraus abgeleiteten Testaufgaben erfolgen. Bei der Entwicklung der Testaufgaben muss beispielsweise darauf geachtet werden, dass die Teilkompetenz (z.B. Leseverstehen) getestet wird, die intendiert ist und sich keine Verzerrung durch andere unerwünschte Aspekte (beispielsweise Weltwissen) ergibt. Sofern die Testaufgaben nicht von Einzelpersonen, sondern in der Gruppe entwickelt werden, können gegenseitige Überprüfungen auf handwerkliche Fehler (z.B. Eignung der Item-Formulierung oder Abhängigkeit von Items untereinander) und die Analyse der für die Lösung notwendigen kognitiven Operationen einer ersten Qualitätssicherung der Aufgaben dienen.

b) *Auswertung der Testaufgaben*
Die Auswertung der Testaufgaben sollte generell möglichst objektiv erfolgen. Bei der Auswertung der Testaufgaben für die rezeptiven Teilkompetenzen werden häufig geschlossene Aufgabenformate verwendet, die in dieser Hinsicht keine Schwierigkeit darstellen, da zur Auswertung Lösungsschlüssel hinterlegt werden. Dennoch sollte bei geschlossenen Aufgaben darauf geachtet werden, dass die Ratewahrscheinlichkeit soweit wie möglich reduziert wird und dementsprechend möglichst keine dichotomen Antwort-Optionen (richtig/falsch, ja/nein) verwendet werden.
Sofern offene Aufgabenformate u.a. für produktive Teilkompetenzen zum Einsatz kommen, muss die mit der Beurteilung einhergehende Subjektivität so gering wie möglich gehalten werden. Als Maßnahmen zur Standardisierung und Objektivierung der Beurteilung sollten analytische und/oder holistische Bewertungskriterien oder andere Bewertungsvorgaben dienen, die von allen Beurteilern einheitlich angewendet und interpretiert werden. Dies sollte durch ein geeignetes Training in der Gruppe und durch Benchmarks, die als Vergleichsgröße für die Bewertungen dienen, sichergestellt werden. Geeignetes Material für Benchmarks steht ggf. durch vorherige Erprobungen zur Verfügung. Durch ein regelmäßiges Monitoring und die Verwendung von Benchmarks wird die Voraussetzung geschaffen, um Unterschiede in Strenge und Milde der Beurteiler zu ermitteln und zu adjustieren, etwa durch Anwendung psychometrischer Modelle wie das Multifacetten-Rasch-Modell (Eckes 2015a, Linacre 1989).

c) *Erprobung von Testaufgaben*
Alle für eine Prüfung oder einen Test vorgesehenen Aufgaben sollten mit einer der Zielgruppe vergleichbaren Probandengruppe erprobt werden. Die Erprobung durch Lehrkräfte allein reicht nicht aus. Gerade bei geschlossenen

Aufgabenformaten lässt sich die Qualität des Aufgabenmaterials ohne empirischen Nachweis bei aller Erfahrung der Testentwickler nicht bestimmen. In der Referenzliteratur wird dabei von zwei Etappen ausgegangen, die als Pilotierung (*piloting*) und Testung oder Normierung (*pretesting* oder *field test*) bezeichnet werden. Bei der Pilotierung geht es darum, das Aufgabenmaterial erstmals auszuprobieren, Unklarheiten oder unverständliche Formulierungen zu entdecken und einen ersten Eindruck von der Schwierigkeit zu gewinnen. Diese erste Etappe kann mit einer kleinen Gruppe von ca. 60-80 Personen durchgeführt werden. Zur psychometrischen Analyse der geschlossenen Aufgaben kann die klassische Item-Analyse herangezogen werden, die die Lösungsraten der Items und Trennschärfen feststellt. Aber auch qualitative Methoden wie Befragungen der Probanden mithilfe von Interviews oder Fragebogen geben Aufschluss über Qualität und Eignung der Aufgaben.
Die zweite Etappe der Testung sollte mit einer größeren Probandengruppe von mindestens 200-300 Personen erfolgen. Das ist vor allem bei standardisierten Tests möglich und im Schulalltag schwer zu realisieren. Diese weitere Erprobung dient der genaueren Feststellung des Schwierigkeitsgrades der Aufgaben unabhängig von der Personenfähigkeit. Dafür eignet sich die Rasch-Analyse (Item-Response-Theorie, IRT), der probabilistische Messmodelle zugrunde liegen (Knoch & McNamara 2015). Ziel ist es, durch die Testung die Items auf einer gemeinsamen Skala abzubilden, d.h. eine Skalierung vorzunehmen (vgl. Kolen & Brennan 2004: 430 *vertical scaling*). Bei diesem Ablauf ist durch geeignete Maßnahmen sicherzustellen, dass der Test die zu messende Fähigkeit zuverlässig und konsistent erfasst und dieser Standard sowie ein vergleichbarer Schwierigkeitsgrad über verschiedene Testversionen und Testläufe hinweg garantiert wird. Dies kann beispielsweise durch die Verwendung von identischen Anker-Items in den verschiedenen Testversionen erreicht werden. Erst wenn diese Voraussetzungen gegeben sind, macht es Sinn, auf der so entwickelten Skala Kompetenzstufen festzulegen. Das heißt, dass die Testwerte der lokalen Skala bestimmt werden, die als Mindestanforderung für das Erreichen einer bestimmten Kompetenzstufe angesehen werden. Der jeweilige Testwert steht im Fall von itembasierten Tests für die Anzahl der Items, die die Testteilnehmenden einer bestimmten Kompetenzstufe richtig gelöst haben müssen. Im Hinblick auf eine kriteriumorientierte Interpretation der Kompetenzstufen werden Beschreibungen der zugrunde liegenden Kompetenzen entwickelt, die über die Leistungsanforderungen Auskunft geben. Sofern ein Bezug zum GeR hergestellt werden soll, können auch diese Kompetenzbeschreibungen verwendet werden.
Die Erprobungsergebnisse von offenen Testaufgaben können durch erfahrene Beurteiler evaluiert werden, die Kommentare zur Bearbeitung der Aufgaben

und beispielsweise Hinweise auf die Verständlichkeit der Aufgabenformulierung abgeben. Die Häufigkeitsverteilung der Testergebnisse und Noten in den offenen Aufgaben geben Hinweise zu der Schwierigkeit der Aufgaben.

Die zuvor unter a) bis c) dargestellten Maßnahmen zum Qualitätsnachweis von Testverfahren verfolgen das Ziel, sowohl die Auswahl von Inhalten und Kompetenzen als auch die Entwicklung des dazu passenden Testmaterials und seine Auswertung soweit als möglich zu standardisieren, damit eine Beeinträchtigung oder Verzerrung der Ergebnisse durch unerwünschte Faktoren wie z.B. Beurteilerstrenge möglichst ausgeschlossen werden und die tatsächlich zu messende Kompetenz im Hinblick auf Zielgruppe und Verwendungszweck auch bei wiederholten Testläufen präzise und gleichbleibend erfasst wird. Die genannten Maßnahmen beinhalten lediglich eine Auswahl, weitere Methoden zur Validierung sind der Referenzliteratur zu entnehmen. Insofern fällt es schwer, die Auswahl weiter einzuschränken oder Kompromisse zuzulassen, insbesondere, wenn es sich um *high-stakes tests* wie bei Abiturprüfungen handelt. Daher sollten möglichst alle unter a) bis c) genannten Maßnahmen abgedeckt werden. Falls Erprobungen mit einer Probandengruppe nicht möglich sein sollten, wäre es denkbar, zumindest vor der Prüfung Gutachten durch Experten erstellen zu lassen und psychometrische Auswertungen nach erfolgter Prüfung durchzuführen, um beispielsweise dysfunktionale Items oder Aufgaben nachträglich von der Bewertung ausschließen zu können bzw. die Bewertung bei produktiven Aufgaben entsprechend anzupassen. Dennoch ist zu bedenken, dass ohne die Verfügbarkeit der zuvor beschriebenen statistischen Auswertungen, insbesondere der Anwendung der IRT, die Mehrheit der im *Manual* aufgeführten Standard-Setting-Methoden (vgl. CoE 2009: Kap. 6; Kaftandjieva 2010) zur Anbindung von Testaufgaben an den GeR oder ein anderes Referenzsystem nicht verwendet werden kann.

4. Verfahren zur Validierung eines GeR-Bezugs

Die Validierung der Niveaustufe(n) einer Prüfung oder eines Tests gilt als Bestandteil der Konstruktvalidierung und repräsentiert einen Kernbereich der Validität. Die Zuordnung von Tests zu einem externen Referenzsystem wie dem GeR kann nur dann als sinnvoll angesehen werden, wenn die Tests bestimmte Verfahren der Standardisierung berücksichtigen, die eine Konstanthaltung der Anforderungen garantieren (vgl. Abschnitt 3). Aus diesem Grund legen auch die Autoren des *Manuals* einen großen Wert auf den Qualitätsnachweis der Sprachtests, die eine Anbindung an eine GeR-Niveaustufe für sich in Anspruch nehmen (vgl. CoE 2009: Kap. 7.2). Dieser Qualitätsnachweis muss demzufolge als erster

Schritt und grundlegende Voraussetzung für eine weitere Validierung des GeR-Bezugs angesehen werden. Geeignete Verfahren für nachgelagerte Validierungsetappen werden im Folgenden erläutert und im Hinblick auf die notwendigen Rahmenbedingungen diskutiert.

Die Einteilung von Kompetenzskalen in verschiedene Niveaustufen erfüllt im Allgemeinen zwei Funktionen: Zum einen wird auf diese Weise eine dichotome Entscheidung über das Bestehen oder Nicht-Bestehen der in einem Testverfahren operationalisierten Anforderungen von zuvor festgelegten Standards gefällt. Sofern es sich bei den Standards um den GeR handelt, würde dies bedeuten, dass eine Prüfung auf einem einzigen GeR-Niveau angesiedelt ist. Zum anderen kann durch eine Einteilung in mehrere Kompetenzstufen eine differenziertere Unterteilung der unterschiedlichen Leistungsstufen dargestellt werden, etwa elementar, mittel, fortgeschritten oder A1, A2, B1, B2, C1, C2 wie im GeR. Übertragen auf Prüfungen würde das bedeuten, dass zwei (z.B. im *Test Deutsch als Fremdsprache* – TestDaF) oder mehrere GeR-Stufen in einer Prüfung berücksichtigt werden. Bei der Einteilung in solche Kategorien werden Leistungen und Fähigkeitswerte (*cut-scores*) in einer Prüfung festgelegt, die als operationalisierter Standard für die jeweilige Kompetenzstufe(n) gelten. Dieses Prozedere kann urteilsbasiert und unter Einbeziehung von Experten ablaufen oder mithilfe von statistischen Methoden vorgenommen werden. In beiden Fällen wird dies als Standard-Setting bezeichnet (Kaftandjieva 2010: 12f.). Sofern diese Einteilung in Leistungskategorien und die Niveaustufen-Zuordnung durch statistische Verfahren vorgenommen werden, erfolgt die Zuordnung zum GeR indirekt durch die Einbindung entsprechender zum GeR kalibrierter Bezugsgrößen und erfordert keine finanziell und personell aufwendige Organisation von urteilsbasierten Verfahren mit Experten-Panels. Urteilsbasierte Verfahren haben den Vorteil einer direkten Zuordnung der Prüfung zum GeR durch die Urteilsfindung der beteiligten Experten, sollten jedoch nach Möglichkeit statistisch ausgewertet werden, um die Qualität der Urteilsfindung (z.B. Konsistenz der Experten, Grad der Übereinstimmung) zu dokumentieren und die Bestehensgrenzen (*cut-off*) zu errechnen. Im *Manual* werden statistische Verfahren und urteilsbasierte Ansätze kombiniert (vgl. CoE 2009).

4.1 Statistische Methoden des Standard-Settings

Wie bereits in Abschnitt 3 dargestellt, werden verschiedene psychometrische Methoden dazu verwendet, die Qualität von Test-Items und Testaufgaben nachzuweisen sowie einen gleichbleibenden Schwierigkeitsgrad des Aufgabenmaterials über verschiedene Testversionen hinweg zu gewährleisten. Abgesehen von dieser

Funktion der Standardisierung können sie auch eingesetzt werden, um Testaufgaben an Referenzsysteme wie den GeR anzubinden. Die folgenden fünf Methoden werden von North (2000: 556-57) als "klassische" Methoden der Verbindung von Messverfahren (*linking*) bezeichnet und beinhalten vier statistische Methoden (a-d) und eine urteilsbasierte (e). Die Methoden unterscheiden sich in der Stärke der Verbindung oder Zuordnung und sind in absteigender Graduierung aufgeführt (vgl. Kecker 2011: Kap. 4.2; Kolen & Brennan 2004):

a) Testangleichung (*equating*),
b) Kalibrierung (*calibrating*),
c) statistische Adjustierung (*statistical moderation*),
d) prognostische Vorhersage (*predicting)*,
e) konsensgeleitete Adjustierung (*social moderation or standards-oriented assessment*).

Die vierte statistische Methode *predicting* wird hier nicht berücksichtigt. Die urteilsbasierte Methode *social moderation* ähnelt dem Verfahren des *Manuals* (CoE 2009: Kap. 5) *standardisation training and benchmarking* und wird unter Abschnitt 4.2 behandelt.

Die Testangleichung (*equating*) bietet aus psychometrischer Sicht die stärkste Verbindung zwischen zwei Bezugsgrößen, konsensgeleitete Adjustierung dagegen die schwächste, da sie nicht auf statistischen Analysen basiert, sondern auf der Urteilsfindung von Personen. Die einzelnen Methoden binden auf unterschiedliche Art und Weise die Bezugsgrößen ein, die den Standard repräsentieren, zu dem eine Verbindung hergestellt werden soll, in diesem Fall der GeR. Diese Bezugsgrößen können in Form von Anker-Tests, Anker-Items (Testangleichung), Leistungsbeispielen (statistische Adjustierung), Skalen (Kalibrierung) oder Bewertungskriterien und standardorientierte Bewertungen von Beurteilern (konsensgeleitete Adjustierung) in die jeweilige Methode integriert werden. Testangleichung und Kalibrierung können mit dichotomen oder polytomen Aufgaben angewendet werden und bedingen den Einsatz von Modellen der IRT. Dabei werden Anker-Items oder Anker-Tests als Eichmaß und GeR-Bezugsgröße eingesetzt, um die Bestehensgrenzen über verschiedene Testversionen hinweg anzupassen und das entsprechende GeR-Niveau zu gewährleisten. Die Anker-Tests (z.B. der DIALANG) oder Anker-Items (vgl. CD-ROM, CoE 2005) müssen jedoch zuvor zum GeR kalibriert worden sein. Mithilfe der Kalibrierung (*calibrating*) können auch Versionen zweier verschiedener Tests mit ähnlichem Konstrukt durch die Einbindung von Anker-Items auf einer gemeinsamen Skala kalibriert und somit statistisch auf einer GeR-Stufe verortet werden. Die Kalibrierung wird darüber hinaus häufig dazu verwendet, Items unterschiedlicher Schwierigkeit auf einer gemeinsamen Skala anzuordnen mit dem Ziel, eine Item-Bank aufzubauen.

Durch die Anwendung der statistischen Adjustierung (*statistical moderation*) können Beurteilungen oder Auswertungen einem Standard angeglichen werden, der durch einen ganzen, bereits zum Bezugssystem kalibrierten Test oder durch einzelne Leistungsbeispiele dargestellt wird, mit dem die Beurteilungen verglichen werden (vgl. North 2000: 564-565). Diese Methode eignet sich insbesondere für die Standardisierung der Beurteilung produktiver Leistungen und wird ggf. nach einem Testereignis angewendet, um zu große Strenge oder Milde von Beurteilern auszugleichen.

4.2 Urteilsbasierte Standard-Setting-Methoden

In einem *high-stakes test* basiert die Festlegung von *cut-offs* in den meisten Fällen auf der vorherigen Datenermittlung durch Experten. Der *cut-score* wird aufgrund dieser Daten errechnet, die eigentliche Festlegung des *cut-offs* steht dann jedoch in der Verantwortung der Testinstitution. Diese kann den rechnerischen *cut-off* so übernehmen oder ggf. modifizieren. Vertreter von Ministerien oder Behörden können zusätzlich in die endgültige Festlegung des *cut-off* eingebunden werden, wenn politische Erwägungen berücksichtigt werden müssen. Dies ist denkbar bei Tests, deren Auswirkungen von politischer Bedeutung sind, wie etwa Tests, die für den Erwerb der Staatsbürgerschaft oder die Zuwanderung in ein Land absolviert werden müssen. In dem oben beschriebenen Prozess des Standard-Setting stellen die festgelegten *cut-scores* eine Operationalisierung der damit verbundenen Kompetenz oder der Niveaubeschreibungen dar.

Kaftandjieva (2010: 158) nennt 62 verschiedene Methoden des Standard-Setting inklusive verschiedener Variationen einzelner Methoden. Die gängigste Kategorisierung bietet dazu Jaeger (1989). Er unterscheidet die Methoden danach, ob die Experten-Urteile sich auf Test-Items beziehen (testzentrierte Methoden) oder auf Personen (personenzentrierte Methoden). Kaftandjieva (2010: 34f.) differenziert weiter nach der Art der Beurteilungsaufgabe, dem Beurteilungsablauf und nach dem Verfahren der *cut-score*-Berechnung. Die Beurteilungsaufgabe der hinzugezogenen Experten richtet sich danach, ob Items, Leistungsbeispiele oder Personen zu beurteilen sind, die Items dichotom oder polytom sind und in welcher Form die Experten ihr Urteil abgeben (z.B. Schätzung von Lösungswahrscheinlichkeiten oder Klassifikation von Personen). Beim Beurteilungsablauf spielt eine Rolle, ob den Experten Feedback zur Verfügung gestellt wird und welcher Art diese Informationen sind (Informationen zum eigenen Beurteilungsverhalten, zu Item-Schwierigkeiten oder zur Auswirkung der ermittelten *cut-scores* auf die Bestehensquote in der Prüfung). Weitere Unterschiede ergeben sich durch die An-

zahl der Beurteilungsrunden und die Beschränkung auf individuelle Entscheidungen oder Gruppenentscheidungen mit Diskussionen. Einen guten Überblick über die am häufigsten verwendeten Standard-Setting-Methoden und die Qualitätsanforderungen an solche Verfahren bieten Cizek (2012), Cizek & Bunch (2007) das *Manual* des Europarats (CoE 2009) und Kaftandjieva (2010).

Die Auswahl einer geeigneten Methode zum Standard-Setting richtet sich zunächst nach der Beschaffenheit des Tests: Dabei ist zu prüfen, ob er lediglich dichotome oder (auch) polytome Items enthält oder zusätzlich offene Aufgaben für mündliche und schriftliche Produktion bzw. Interaktion. Des Weiteren sollte berücksichtigt werden, dass die Beurteileraufgabe in einigen Methoden ein relativ abstraktes Konzept beinhaltet, das vielen Experten Schwierigkeiten bei der Umsetzung bereitet. Beispielsweise die Vorstellung, ob ein mindestkompetenter Lerner der GeR-Stufe B1 mit einer Wahrscheinlichkeit von 67% ein vorliegendes Item im Leseverstehen richtig lösen würde oder nicht (vgl. Angoff-Methode oder Bookmark-Methode). Hinweise zu ähnlichen und anderen Problemen, die mit der Durchführung von Standard-Setting-Verfahren verbunden sind, finden sich in empirischen Untersuchungen von Figueras & Noijons (2009), Kecker (2011) und Martyniuk (2010) zur Pilotierung des *Manuals* und den dabei eingesetzten Standard-Setting-Methoden. Porsch, Tesch & Köller (2010a) beziehen sich in ihren Anmerkungen dazu auf die Standard-Setting-Methoden, die zur Implementierung der Bildungsstandards für Französisch angewendet wurden.[4] Vor diesem Hintergrund scheint es ratsam, im Allgemeinen eher Methoden anzuwenden, die ein holistisches Urteil über Item-Beispiele (rezeptive Teilkompetenzen) oder Leistungsbeispiele (produktive Teilkompetenzen) von den Experten verlangen. Zu diesen Methoden gehören u.a. die Body-of-Work-Methode (Kingston & Tiemann 2012; Cizek & Bunch 2007: 123f.; CoE 2009: 70) und das Benchmarking (CoE 2009: Kap. 5; North & Jones 2009: 15f.).

Bei der Anwendung der Body-of-Work-Methode beurteilen Experten verschiedene heterogene Leistungsbeispiele von Testkandidaten in Form eines Dossiers (Aufsätze, Antworten zu geschlossenen oder halboffenen Items). Ziel ist es, die Ergebnisse dieser verschiedenen Aufgabenformate als Gesamtleistung zu beurteilen und einer Kompetenzstufe zuzuweisen. Das Verfahren setzt eine numerische Gesamtnote aller Einzelleistungen voraus. Die Experten erhalten in den Dossiers Leistungssets zu einer bestimmten Anzahl von Testteilnehmenden und müssen diese den vorgegebenen Kompetenzstufen zuordnen. Das Vorgehen erfolgt in zwei Schritten, einer Grobeinteilung (*rangefinding*) und einer Feineinteilung (*pinpointing*), die sich auf Leistungssets aus dem Grenzbereich konzentriert. Bei der Feineinteilung werden Leistungsbeispiele, die eindeutig einer Kompetenzstufe

4 Vgl. auch den Beitrag von Neil Jones in diesem Heft.

zuzuordnen sind, bereits nicht mehr berücksichtigt, sondern es werden zusätzliche Beispiele aus dem Grenzbereich hinzugefügt, die zuzuordnen sind. Auf der Grundlage der Feineinstufung wird schließlich der *cut-off* mittels logistischer Regression (vgl. CoE 2009: Kap. 6.6.2) ermittelt.

Für das Benchmarking werden als Erstes Leistungsbeispiele zur Illustration der Referenzskala ausgewählt, die als repräsentativ für die jeweilige Kompetenzstufe gelten können. Solche Beispiele wurden während des Niveaustufen-Projekts des Europarats in Benchmarking-Konferenzen zusammengestellt sowie von Testanbietern zur Verfügung gestellt (für die rezeptiven Teilkompetenzen: CD-ROM, CoE 2005; für den mündlichen Ausdruck: Bolton, Glaboniat, Lorenz, Perlmann-Balme, & Steiner 2008; Breton, Lepage & North 2008; für den schriftlichen Ausdruck: www.coe.int/lang). Anhand dieser vorausgewählten Beispiele mit bekannter Niveaustufe werden Experten in der Beurteilung der Beispiele unter Verwendung der Referenzskala trainiert. Auf diese Weise soll ein einheitliches Verständnis der Kompetenz erworben werden, die zu der Leistung auf einer bestimmten Niveaustufe der Skala gehört. Diese standardorientierte Beurteilung wird in einem weiteren Schritt auf die Beurteilung lokaler Leistungsbeispiele aus dem Test übertragen, der zum GeR zugeordnet werden soll. Das Verfahren findet in mehreren Beurteilungsrunden statt, in denen sowohl die Konsistenz der Beurteilung als auch die Übereinstimmung der Experten untereinander dokumentiert und ausgewertet werden. Je nach Entscheidung der Organisatoren kann Feedback und Diskussion der Beurteilung in die Urteilsfindung einfließen. Eine genaue detaillierte Anleitung zum Benchmarking findet sich im *Manual* des Europarats (CoE 2009: Kap. 5).

Eine Variation dieser Methode ist die konsensgeleitete Adjustierung (vgl. North 2000: 557), die als urteilsbasiertes Verfahren häufig verwendet wird, um die Beurteilung produktiver Leistungen durch Lehrkräfte stärker zu vereinheitlichen. Der Fokus liegt bei der konsensgeleiteten Adjustierung stärker auf einer Kompetenzentwicklung von Beurteilern (standardorientiertes Training) als auf der Niveaustufen-Zuordnung einer Prüfung oder eines Tests. Das heißt, dass die im Training erworbene Kompetenz auf zukünftige Beurteilungen angewendet wird, jedoch nicht auf die GeR-Zuordnung von lokalen Testaufgaben und Leistungsbeispielen wie beim Benchmarking. Konsensgeleitete Adjustierung wird daher hier nicht den Standard-Setting-Methoden zugerechnet, obwohl sie genau wie diese ein urteilsbasiertes Verfahren darstellt. Im *Manual* werden konsensgeleitete Moderation und Benchmarking als *standardisation training and benchmarking* bezeichnet (vgl. CoE 2009: Kap. 5).

4.3 Methodischer Ansatz des *Manuals*

Die Autoren des *Manuals* haben eine Reihe der zuvor dargestellten statistischen und urteilsbasierten Verfahren der Standardisierung und Niveaustufen-Zuordnung zu einem eigenen methodischen Ansatz kombiniert. Zusätzlich wurden Phasen integriert, die der Vertrautmachung mit dem Referenzsystem GeR und der Niveaueinstufung von Items und Leistungsbeispielen auf den GeR-Skalen dienen. Auf diese Weise ergeben sich die folgenden fünf Phasen:

1) *familiarisation* (Vertrautmachung mit dem GeR),
2) *specification* (Beschreibung des Testformats mit dem deskriptiven System und den Skalen des GeR),
3) *standardisation training and benchmarking* (Training und Standardisierung von Leistungsbeurteilung mit Bezug zum GeR),
4) *standard setting* (Festlegung von Bestehensgrenzen durch Urteilsfindung von Experten),
5) *empirical validation* (Qualitätsnachweis der Prüfung oder des Tests, Qualitätsnachweis des Standards-Setting, Überprüfung der ermittelten Bestehensgrenze durch alternative Verfahren, z.B. Korrelation mit anderen Messverfahren).

Die Phasen *specification* und *standardisation/benchmarking* sind rein urteilsbasiert und erfordern eine statistische Auswertung lediglich zur Evaluierung des Verfahrens. Die einzelnen Phasen werden detailliert im *Manual* beschrieben und durch zusätzliche Kapitel im Anhang zu Forschungsmethoden, zum Standard-Setting und zum Einsatz psychometrischer Analysen ergänzt (sog. *Reference Supplement*, Takala 2009).

Die zuvor beschriebenen Methoden zum Qualitätsnachweis eines Tests, zur Standardisierung und zur Niveaustufen-Zuordnung erfordern fast alle den Einsatz von statistischen Verfahren und psychometrischen Analysen. Ohne diese gilt kein *high-stakes test* als qualitativ hochwertig und entspricht nicht den international für Testverfahren im Bildungsbereich zugrunde gelegten Standards (vgl. AERA, APA & NCME 2014).

Sofern sich Bildungseinrichtungen dazu entscheiden, beispielsweise den methodischen Ansatz des *Manuals* umzusetzen, jedoch ohne statistische Verfahren anzuwenden, so sollten zumindest die ersten drei genannten Phasen durchgeführt werden. Sie dienen der Familiarisierung mit dem GeR, der standardisierten Beschreibung der Prüfung oder des Tests und der damit verbundenen möglichst einheitlichen Erstellung der dazugehörigen Items und Aufgaben sowie einer standardorientierten Beurteilung der produktiven Leistungen. Der in der fünften Phase angesiedelte Qualitätsnachweis für Sprachprüfungen sollte ebenfalls in

geeigneter Form erbracht werden (vgl. Abschnitt 3). Darüber hinaus können urteilsbasierte Methoden des Standard-Setting wie Benchmarking oder Body-of-Work-Methode, die eher holistisch vorgehen, verwendet werden, um Bestehensgrenzen festzulegen und eine erste Zuordnung zu GeR-Niveaustufen vorzunehmen.

5. Schlussbemerkungen

Die Einführung von PISA-Studien auf internationaler und auch nationaler Ebene hat in Europa und auch in Deutschland ein größeres Bewusstsein für Qualitätsstandards im Testen und Prüfen von Kompetenzen geschaffen. Ein weiterer Schritt in diese Richtung wurde durch die Pilotierung des *Manuals* im Rahmen des Europarat-Projekts *Relating Language Examinations to the Common European Framework of Reference* (2004-2008) unternommen. Methoden wie das Standard-Setting zur Festlegung von Niveaugrenzen auf Kompetenzskalen wurden zwar im Rahmen der nationalen PISA-Studien in Deutschland bereits umgesetzt, jedoch von europäischen und deutschen Sprachtestanbietern erstmalig im Zusammenhang mit dem Europarat-Projekt durchgeführt. Zu nennen sind auf europäischer Ebene beispielsweise der computerbasierte diagnostische Online-Test DIALANG (vgl. Alderson 2005, http://www.lancaster.ac.uk/researchenterprise/dialang/about) und das europäische Projekt *European Survey on Language Competences* (ESLC) (European Commission 2012) zur Erfassung der Fremdsprachenkenntnisse von Schülerinnen und Schülern am Ende der Sekundarstufe I, die 2011 in 14 europäischen Ländern erhoben wurden. Darüber hinaus haben Mitglieder der ALTE, darunter das Goethe-Institut (vgl. für das B1-Zertifikat www.goethe.de) oder das TestDaF-Institut (vgl. Kecker 2011; Kecker & Eckes 2010), aber auch deutsche Bildungseinrichtungen wie das IQB (vgl. Porsch, Tesch & Köller 2010a), verschiedene Methoden des Standard-Setting eingesetzt, um den Bezug ihrer Sprachprüfungen oder von Testaufgaben zu den jeweiligen GeR-Niveaustufen nachzuweisen. Die Frage stellt sich jedoch, wie dieser Kompetenzzuwachs dem Sektor Schule systematisch zur Verfügung gestellt werden kann, um Testentwicklungen in den Fremdsprachen für Schulabschlussprüfungen zu professionalisieren. Porsch, Tesch & Köller (2010b) haben dargestellt, inwieweit Lehrkräfte aus den verschiedenen Bundesländern an der Entwicklung der Testaufgaben für die Implementierung der Bildungsstandards für den Mittleren Schulabschluss in den Fremdsprachen beteiligt waren. In Zusammenhang mit der Durchführung des Standard-Setting im Fach Französisch schlugen sie vor, ähnliche Veranstaltungen systematischer als bisher in die Weiterbildung von Lehrkräften zu integrieren:

> Die Rückmeldung der Teilnehmer zum Standard-Setting zeigten überdies, dass die Teilnehmer diesen Workshop ausnahmslos als eine persönliche Bereicherung verstanden. Auch im Hinblick auf die Inhalte in den jeweils dreitägigen Veranstaltungen – die Vertrautmachung bzw. Rezeption der Bildungsstandards, die Einführung in den Bereich der Forschung zu schwierigkeitsgenerierenden Merkmalen von Testaufgaben, die Beschreibung der Studienkonzeption und -durchführung, der Diskussion über die Konstruktion von guten Testaufgaben und ausstehende empirische wie fachdidaktische Arbeiten im Fach Französisch – besitzt ein solches Verfahren Modellcharakter als Instrument der schulgestützten Lehrerbildung. Es wäre durchaus vorstellbar, Fachkollegien an Schulen mit empirisch ermittelten Schwierigkeiten von Testaufgaben in Fortbildungen zu konfrontieren, um fachinterne Diskussion über aufgabenbezogenes Lernen zu beflügeln und fachspezifisches Wissen etwa zu den schwierigkeitsgenerierenden Merkmalen von Hör- und Leseverstehensaufgaben zu vermitteln (Porsch, Tesch & Köller 2010b: 265).

Vor kurzem wurde vom IQB mit Beteiligung von Lehrkräften und Fachdidaktikern auch eine Aufgabensammlung für das Zentralabitur in den modernen Fremdsprachen entwickelt, die der Einführung der Bildungsstandards für die Allgemeine Hochschulreife dient, die 2012 von der KMK verabschiedet wurden. Es liegt nunmehr in der Hand der Bildungsbehörden auf Länderebene, das Wissen der auf diese Weise geschulten Lehrkräfte für die Weiterbildung in ihren Bildungseinrichtungen zu nutzen. Abgesehen von solchen Maßnahmen der Weiterbildung fehlt jedoch bereits in der Lehrerausbildung sowohl im Studium als auch im Referendariat die Vermittlung von Fachkompetenz in einigen Bereichen der Leistungsmessung für den schulischen Bedarf. In dieser Hinsicht bleibt noch einiges zu tun, um Lehrer hinreichend auf alle Facetten ihres Berufs vorzubereiten.

Eingang des revidierten Manuskripts 15.12.2015

Literaturverzeichnis

AERA, APA & NCME = American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014), *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alderson, J. Charles (Hrsg.) (2002), *Common European Framework of Reference for Languages: Learning, teaching, assessment − case studies*. Strasbourg: Council of Europe.

Alderson, J. Charles (2005), *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. Charles (2007), The challenge of (diagnostic) testing: Do we know what we are measuring? In: Fox, Janna; Wesche, Marjorie; Bayliss, Doreen; Cheng, Liying; Turner, Carolyn E. & Doe, Christine (Hrsg.) (2007), *Language testing reconsidered*. Ottawa: University of Ottawa Press, 21-39.

Alderson, J. Charles; Figueras, Neus; Kuijper, Henk; Nold, Günther; 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* [Online: http://eprints.lancs.ac.uk/44/1/final_report.pdf, 02.11.2015].

Alderson, J. Charles; Figueras, Neus; Kuijper, Henk; Nold, Günther; 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 & Palmer, Adrian (2010), *Language Assessment in Practice*. Oxford: Oxford University Press.

BIFIE (Hrsg.) (2012), *Bildungsstandards in Österreich. Überprüfung und Rückmeldung* (4. aktualisierte Aufl.). Salzburg.

Bolton, Sibylle; Glaboniat, Manuela; Lorenz, Helga; Perlmann-Balme, Michaela & Steiner, Stefanie (2008), *Mündlich. Mündliche Produktion und Interaktion Deutsch. Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens für Sprachen*. Berlin: Langenscheidt.

Breton, Gilles; Lepage, Sylvie & North, Brian (2008), *Cross-language benchmarking seminar to calibrate examples of spoken production in English, French, German, Italian and Spanish with regard of the six levels of the Common European Framework of Reference for Languages (CEFR)* [Online: http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Sevres_Report_2008_EN.pdf, 13.01.2016].

Cizek, Gregory J. (Hrsg.) (2012), *Setting Performance Standards. Foundations, Methods, Innovations* (2. Aufl.). New York: Routledge.

Cizek, Gregory. J. & Bunch, Michael B. (2007), *Standard setting. A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

CoE = Council of Europe (2001), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.

CoE = Council of Europe (2005), *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Reading and listening items and tasks: Pilot samples illustrating the common reference levels in English, French, German, Italian and Spanish*. Strasbourg: Language Policy Division.

CoE = 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 [Online: http://www.coe.int/t/dg4/linguistic/Source/Manual Revision-proofread-FINAL_en.pdf, 13.01.2016].

Davies, Alan; Brown, Annie; Elder, Cathie; Hill, Kathryn; Lumley, Tom & McNamara, Tim (1999), *Dictionary of language testing*. Cambridge: Cambridge University Press.

Downing, Steven M. & Haladyna, Thomas M. (Hrsg.) (2006), *Handbook of test development*. Mahwah, NJ: Erlbaum.

Eckes, Thomas (2015a), *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2. Aufl.). Frankfurt a.M.: Peter Lang.

Eckes, Thomas (2015b), Validität: Flexionen eines polymorphen Konzepts. In: Böcker, Jessica & Stauch, Anette (Hrsg.) (2015), *Konzepte aus der Sprachlehrforschung – Impulse für die Praxis. Festschrift für Karin Kleppin*. Frankfurt a.M.: Peter Lang, 449-468.

Europarat (2001), *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

European Commission (2012), *First european survey on language competences: Technical repor*t. Strasbourg [Online: http://ec.europa.eu/languages/policy/strategic-framework/documents/language-survey-technical-report_en.pdf, 02.11.2015].

Figueras, Neus & Noijons, José (Hrsg.) (2009), *Linking to the CEFR levels: Research perspectives.* Arnhem: Cito, EALTA.

Fulcher, Glenn & Davidson, Fred (2007), *Language testing and assessment, an advanced resource book*. London/New York: Routledge.

Galaczi, Evelina D. & Weir, Cyril (Hrsg.) (2013), *Exploring Language Frameworks*. *Proceedings of the ALTE Kraków Conference July 2011*. Studies in Language Testing, 36. Cambridge: CUP.

Grotjahn, Rüdiger (2003), *Leistungsmessung und Leistungsbewertung, Studienbrief, Erprobungsfassung 12/2002*, Hagen: FernUniversität Hagen.

Harsch, Claudia (2005), *Der Gemeinsame europäische Referenzrahmen für Sprachen*: *Leistung und Grenzen. Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt. Inaugural-Dissertation*. Universität Augsburg.

Harsch, Claudia (2014), General language proficiency revisited: current and future issues. *Language Assessment Quarterly* 11: 2, 152-169.

Harsch, Claudia (2015), Assessment Literacy. In: Böcker, Jessica & Stauch, Anette (Hrsg.) (2015), *Konzepte aus der Sprachlehrforschung – Impulse für die Praxis. Festschrift für Karin Kleppin*. Frankfurt a.M.: Peter Lang, 489-509.

Hulstijn, Jan H. (2007), The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91: 4, 663-667.

Jaeger, Richard M. (1989), Certification of student competence, In: Linn, Robert L. (Hrsg.) (1989), *Educational measurement* (3. Aufl.). Washington, DC: American Council on Education/ Macmillan, 485-514.

Jones, Neil (2013), Defining an inclusive framework for languages. In: Galaczi, Evelina D. & Weir, Cyril (Hrsg.) (2013), 105-117.

Kaftandjieva, Felianka (2010), *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to test reading in EFL*. Arnhem: Cito.

Kane, Michael (2006), Validation. In Brennan, Robert L. (Hrsg.) (2006), *Educational measurement* (4. Aufl.). New York: American Council on Education/Praeger, 17-64.

Kane, Michael (2012), Validating score interpretations and uses. *Language Testing* 29: 1, 3-17.

Kecker, Gabriele (2011), *Validierung von Sprachprüfungen: Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen*. Frankfurt a.M.: Peter Lang.

Kecker, Gabriele, & Eckes, Thomas (2010), Putting the Manual to the test: The TestDaF–CEFR linking project. In: Martyniuk, Waldemar (Hrsg.) (2010), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press, 50-79.

Kingston, Neal M. & Tiemann, Gail C. (2012), *Setting Performance Standards on Complex Assessments. The Body of Work Method.* In: Cizek, Gregory (Hrsg.) (2012), 201-223.

Kleppin, Karin (2003), Der Gemeinsame europäische Referenzrahmen für Sprachen: Ärgernis oder Fortschritt? In: Bausch, Karl-Richard; Christ, Herbert; Königs, Frank & Krumm, Hans-Jürgen (Hrsg.), *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*. Tübingen: Narr, 105-112.

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2003), *Bildungsstandards für die erste Fremdsprache (Englisch/ Französisch) für den Mittleren Schulabschluss*. *Beschluss der Kultusministerkonferenz vom 04.12.2003* [Online: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/ 2003_12_04-BS-erste-Fremdsprache.pdf, 05.02.2016].

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2012), *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch für die allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012* [Online: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Fortgef-FS-Abi.pdf, 13.01.2016].

Knoch, Ute & McNamara, Tim (2015), Rasch analysis. In: Plonsky, Luke (Hrsg.) (2015), *Advancing quantitative methods in second language research*. New York: Routledge, 275-304.

Kolen, Michael J. & Brennan, Robert L. (2004), *Test equating, scaling, and linking: Methods and practices* (2. Aufl.). New York: Springer.

Konsortium HarmoS Fremdsprachen; Schneider, Günther; Lenz, Peter & Studer, Thomas (Hrsg.) (2009), *Fremdsprachen. Wissenschaftlicher Kurzbericht und Kompetenzmodell.* Bern: EDK [Online: http://www.edudoc.ch/static/web/arbeiten/harmos/L2_wissB_25_1_10_d.pdf, 28.09.2015].

Linacre, John M. (1989), *Many-facet Rasch measurement*. Chicago: MESA Press.

Martyniuk, Waldemar (Hrsg.) (2010), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual*. Cambridge, UK: Cambridge University Press.

Messick, Samual (1989), Validity. In: Linn, Robert L. (Hrsg.) (1989), *Educational measurement* (3. Aufl.). New York: Macmillan, 13-103.

Milanovic, Michael (2009), Cambridge ESOL and the CEFR. *Research Notes* 37, 2-5.

North, Brian (2000), Linking language assessments: An example in a low stakes context. *System* 28: 4, 555-577.

North, Brian & Jones, Neil (2009), *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. Strasbourg: Council of Europe/Language Policy Division.

Porsch, Raphaela; Tesch, Bernd & Köller, Olaf (Hrsg.) (2010a), *Standardbasierte Testentwicklung und Leistungsmessung*. Münster: Waxmann.

Porsch, Raphaela; Tesch, Bernd & Köller, Olaf (2010b), Die Entwicklung von Kompetenzstufenmodellen zum Lese- und Hörverstehen im Fach Französisch. In: Porsch, Raphaela; Tesch, Bernd & Köller, Olaf (Hrsg.) (2010a), 244-266.

Schneider, Günther & North, Brian (2000), *Fremdsprachen können - was heißt das?* Chur/ Zürich: Rüegger.

Takala, Sauli (Hrsg.) (2009), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe/Language Policy Division.

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

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 des Deutschen*. Frankfurt a.M.: Peter Lang.