Vorwort zum Themenheft "Kompetenzorientiert testen und prüfen"

Testen und Prüfen gehören zum Lehren und Lernen von Fremdsprachen; sie strukturieren alle Lernanstrengungen durch eine Beurteilung von Lernerfolgen und sprachlichen Kompetenzen und legen die Grundlage für eine zielgerichtete Diagnose und darauf aufbauende weiterführende Entscheidungen – etwa über die Zulassung zu einem Studium, die Einstufung in einen Sprachkurs, einen Abschluss in einem Sprachenfach des schulischen Bildungssystems oder auch den Verbleib in einem Zuwanderungsland. Testen und Prüfen haben deshalb einen sehr hohen Stellenwert für Individuen, die Sprachen erlernen wollen oder auch müssen, denn ihre Konsequenzen sind oft weitreichender als die unmittelbare Prüfung. Im Sprachgebrauch der Testdidaktik nennt man solche Prüfungen auch *high-stakes tests*. Oft hängt es von einem bestimmten Schwellenwert (*cut-off point*) ab, ob ein Ziel erreicht wurde oder nicht. Diese Erkenntnis hat zur Folge, dass Testen und Prüfen von Testanbietern im Bereich von Sprachzertifikaten und von Lehrkräften im schulischen Bildungskontext mit größter Sorgfalt und hohem Verantwortungsbewusstsein durchgeführt werden müssen (vgl. auch Rossa 2016).

Die Begriffe "Testen" und "Prüfen" beziehen sich auf unterschiedliche Formen dessen, was man heute zusammenfassend als "Assessment" (Beurteilung) bezeichnet, wobei mit "Testen" in der Regel die stärker standardisierten Formen der Lernstandsbeurteilung angesprochen werden. Tests und Prüfungen bestehen in der Regel aus mehreren Prüfungsteilen, können also auch mehrere Beurteilungsformate umfassen. Lernbegleitende Prüfungen (*formative assessement*), die eher einer Diagnose des Lernprozesses dienen und eine regelmäßige Rückmeldefunktion für die Lernenden haben, werden wiederum unterschieden von punktuellen Beurteilungen der Kompetenz, die oftmals eine Lernsequenz abschließen (*summative assessment*). Im vorliegenden Themenheft wird es vor allem um die theoretischen Grundlagen kompetenzorientierter summativer Tests und Prüfungen gehen. Dabei sollen sowohl Prüfungen im schulischen Bereich als auch Tests von professionellen Testanbietern berücksichtigt werden.

Im zurückliegenden Jahrhundert wurden Prüfungen im Fremdsprachenunterricht und insbesondere die Testentwicklung von unterschiedlichen, vor allem linguistischen Theorien, aber auch von der Spracherwerbsforschung beeinflusst (vgl. Arras & Kecker 2016). Im traditionellen, strukturalistisch geprägten Ansatz, der von streng gegliederten Teilbereichen des Sprachsystems ausging, wurden auch psychometrische Verfahren zur Messung von einzelnen Kompetenzen herangezogen, bevor eine geänderte Auffassung von Sprachkompetenz eher an der lebensweltlichen kommunikativen Sprachverwendung orientierte, integrative

Messverfahren einführte (vgl. McNamara 1998). Im Zuge der Ausrichtung auf kommunikative Kompetenz änderten sich die Prüfungs- und Testverfahren; mündliche Testformate wurden häufiger, Aufgaben werden heute vorwiegend in kommunikative Situationen eingebettet.

Die Testforschung erfolgte bislang vor allem in angelsächsischen Ländern. Gründe hierfür liegen zum einen im wichtigen internationalen Stellenwert der englischen Sprache, zum anderen in großen Migrationsbewegungen des 20. Jahrhunderts. In Großbritannien, ehemaligen Commonwealth-Staaten, vor allem Australien, Kanada sowie den USA, erlangten Tests und Sprachenzertifikate sehr viel eher als im übrigen Europa große Bedeutung. Die wichtigsten Fachzeitschriften *Language Testing* und *Language Assessment Quarterly* sind ebenfalls im angelsächsischen Raum angesiedelt (vgl. Spolsky 1995; Perlmann-Balme 2016).

Einen Meilenstein stellt die Veröffentlichung des *Gemeinsamen europäischen Referenzrahmens für Sprachen* (GeR, Europarat 2001) dar, der die geschilderte Entwicklung weltweit, vor allem aber im europäischen Raum weiter vorantrieb. Das Dokument veranschaulicht mit seinen Deskriptoren, wie ein standardbasierter Rahmen für moderne kompetenzorientierte Prüfungen aussehen könnte. Seine Impulse werden in vielen Ländern von der Bildungspolitik aufgenommen. Letztlich ist es auch auf dieses Dokument zurückzuführen, dass die großen internationalen Testanbieter in den letzten 15 Jahren eine verstärkte Qualitätskontrolle verbunden mit mehr Anstrengungen im Bereich der Sprachtestforschung unternommen haben (Arras & Kecker 2016).

Im deutschen Bildungssystem begann die Kultusministerkonferenz (KMK) nach dem PISA-Schock auf der Grundlage des GeR Bildungsstandards zu entwickeln, zunächst für den Englisch- und Französischunterricht für den mittleren Schulabschluss bzw. für die Hauptschule (KMK 2004, 2005). Weil die Referenzniveaus, die der GeR vorschlug, kompetenzorientiert waren, steuerte man im Bildungssystem auch die Prüfungen für den Hauptschulabschluss und den mittleren Bildungsabschluss in diese Richtung. Damit wurde im Grunde etwas eingelöst, das ohnehin für den Fremdsprachenunterricht seit den 1970er Jahren immer wieder gefordert worden war, ein am Lernziel "Kommunikative Kompetenz" (Piepho 1974, van Ek 1975) orientierter Unterricht, dessen Ergebnisse auch mit kommunikativen Beurteilungsverfahren überprüft werden müssen. Der Wechsel zu einem standardorientierten kompetenzorientierten Lehren, Lernen und Beurteilen brachte jedoch gerade in den ersten Jahren nach Erscheinen der Bildungsstandards viel grundlegende Kritik von Seiten der Sprachlehrforschung und Fremdsprachendidaktik hervor (vgl. Bausch, Christ, Königs & Krumm 2003 und Bausch, Burwitz-Melzer, Königs & Krumm 2005). Dabei stand oft im Mittelpunkt der Einwände, dass literarisch-ästhetische Kompetenzen in den Standards ausgeblendet und zentrale Zielvorstellungen des Fremdsprachenunterrichts wie

die Persönlichkeitsbildung nicht abgebildet worden waren (Burwitz-Melzer 2005: 57-61). Die DGFF veröffentlichte in der ZFF (Caspari, Grünewald, Hu, Küster, Nold, Vollmer & Zydatiß 2008) eine Stellungnahme, die für eine Offenheit gegenüber den bildungspolitischen Neuerungen plädierte. 2012 wurden nach punktueller empirischer Erprobung die *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife* veröffentlicht; sie sollen auf ein Zentralabitur in den Hauptfächern Deutsch, Mathematik und den fortgeführten Fremdsprachen ab 2017 vorbereiten. Die ebenfalls am GeR orientierten, jedoch sorgfältiger auf die Lehr- und Bildungsinhalte abgestimmten Standards wurden insgesamt positiver aufgenommen. Dazu mögen auch die beigefügten Sammlungen von Beispielen an Lern- und Prüfungsaufgaben beigetragen haben. Die Abiturprüfungen, die in diesem Dokument auch geregelt werden, sind nun standardorientiert angelegt. Wie erfolgreich der Paradigmenwechsel zum standard- und kompetenzorientierten Lernen, Lehren und Prüfen an deutschen Schulen sein wird, steht heute noch nicht fest. Hier liegt ein Forschungsdesiderat für die unmittelbare Zukunft, denn erst umfassende empirische Forschungsprojekte werden zeigen können, wie erfolgreich Lehrende und Lernende mit den Bildungsstandards umgehen.

Auch außerhalb des deutschen Schulsystems gewann der Einfluss des GeR schnell an Boden: Die übersichtlichen Skalen und vermeintlich erschöpfenden Deskriptoren überzeugten nicht nur Schulbuch-Verlage, sondern auch die Verantwortlichen für die Erwachsenenbildung und deutsche wie europäische Testanbieter. Allerdings war der GeR auch hier als Bezugspunkt und Basis für die Validierung von Tests ein zum Teil problematisches Instrument, weil seine Deskriptoren für Kompetenzen zwar in aufwändigen empirischen Verfahren entstanden, aber gerade dadurch für Benutzer kaum durchschaubar waren (vgl. Alderson, Figueras, Kuijper, Nold, Takala & Tardieu 2004; Harsch 2005). Quetz und Vogt (2009) warnten vor einer vorschnellen Übernahme des GeR: Sie bemängelten u.a., dass die Skalen mit Deskriptoren für Kompetenzniveaus aus heterogenen Quellen abgeleitet wurden und erst später sukzessive in konsensorientierten Verfahren empirisch geprüft wurden. Die vorschnelle Popularisierung der Referenzniveaus in Lehrmaterialien und Tests könnte, so die Autoren, zu einer Trivialisierung führen. Diese vor allem auf theoretische Gesichtspunkte konzentrierte Unsicherheit gegenüber dem GeR führte letztendlich dazu, dass bei den großen Testanbietern in Europa jeweils eigene Testspezifikationen benutzt werden, die zwar auf dem GeR basieren, aber dessen Deskriptoren eigenständig umsetzen. Auch heute gibt es noch zahlreiche offene Fragen zur Umsetzung des GeR in Tests und Prüfungen, die wohl erst mit einer Überarbeitung dieses Dokuments besser beantwortet werden können.

Besonders viele Fragen zielen auf den Aspekt der Fairness von Tests, aber auch auf Fragen der Validität, insbesondere der Konstruktvalidität und des damit zusammenhängenden *standard setting*. Alle drei Aspekte hängen eng zusammen: Seit den 1990er Jahren gibt es eine lebhafte Debatte über die Fairness von Sprachtests. Ein *Language Testing Research Colloquium* befasste sich 1997 mit dem Thema "Fairness in Language Testing" (Davies 1997). Eine faire Beurteilung setzt Validität voraus, die gemeinhin so definiert wird, dass ein Test misst, was er zu messen beabsichtigt: "In order to justify a particular score interpretation, we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and very little else" (Bachman & Palmer 1996: 21). Der Beitrag von **Gabriele Kecker** (Bochum) setzt sich gerade mit der Problematik auseinander, welche Erwartungen man an den GeR in Hinblick auf Sprachprüfungen und -tests richten kann: Welche Kriterien muss ein Test erfüllen, um im Sinne des GeR valide zu sein? Diese Frage spielt vor allem bei den zentralen Abschlussprüfungen der Bundesländer eine Rolle, aber auch in Hinblick auf das für 2017 geplante Zentralabitur. Wir haben Keckers Beitrag an den Anfang dieses Themenhefts gestellt, weil er einen umfassenden Überblick darüber gibt, welche Grundsätze heute bei der Erstellung von *high-stakes tests* beachtet werden müssen.

Grundlage vieler *standard setting*- und Validierungsverfahren ist für die in der ALTE[1] zusammengeschlossenen großen Anbieter von Sprachtests der GeR. Spiegeln die Referenzniveaus und vor allem ihre Deskriptoren das konsensuelle Verständnis von Testforschern und Testkonstrukteuren wider? In seinem Beitrag schildert **Neil Jones** (Cambridge) seine Erfahrungen als Leiter zweier großer europäischer Projekte, dem European Survey on Language Competences (ESLC) und der Study on Comparability of Language Testing in Europe (SCLTE), deren Ergebnisse im September 2015 unter dem Titel "Constructing comparable standards of communicative language competence" veröffentlicht wurden. In diesem Beitrag geht es vor allem um *standard setting*, einem nach wie vor kontroversen Kapitel auch beim Assessment, das man vereinfacht in der Formel beschreiben kann "Is my B1 also your B1?".

Im Kontext der Überlegungen zur Konstruktvalidierung zeigt der Beitrag von **Olaf Bärenfänger** (Leipzig), welche Probleme sich bei der Nutzung der Kompetenzskalen des GeR für die Bewertung schriftlicher Lernertexte in der Praxis ergeben können. Die Texte stammen aus Prüfungen der telc GmbH in den Sprachen Deutsch und Italienisch, die von je zwei Beurteilern für jede Sprache auf die Referenzniveaus eingestuft wurden. Analysiert wird eine große Zahl von Lernertexten. Bärenfänger führt auf der Grundlage der Beurteilungen und Einstufungen eine Multifacetten-Rasch-Analyse durch, die wesentliche Erkenntnisse

1 ALTE = Association of Language Testers in Europe, http://www.alte.org

zur Qualität einiger Skalen des GeR erlaubt. Seine gründlichen statistischen Analysen zeigen, auf wie unsicherem Terrain man sich bei einer Übernahme von Deskriptoren als Beurteilungsgrundlage bewegt, und dass die Diskussion um die Referenzniveaus noch lange nicht als erledigt betrachtet werden kann.

Einen anderen, nicht minder interessanten Aspekt zum Thema Konstruktvalidierung diskutiert **Karin Vogt** (Heidelberg) mit ihrem Aufsatz zu dem hochkomplexen Kompetenzbereich des interkulturellen kommunikativen Lernens. Unsicherheiten bei der Operationalisierung und Validierung dieses Kompetenzbereichs, der für schulische Bewertungen, aber auch in der Erwachsenenbildung eine wichtige Rolle spielt, sind nicht neu: In der Veröffentlichung seines Konzepts der *Communicative Competence* hatte Byram (1997) dazu geraten, diesen Kompetenzbereich aus dem Kontext von Beurteilungen herauszunehmen und neue Formen wie Selbstbeurteilung für die Diagnose eines Lernfortschritts zu nutzen. Dass die Bildungsstandards von 2012 nach der harschen Kritik an den interkulturellen Standards von 2004 und 2005 einen erneuten, weitaus komplexeren Versuch gestartet haben, diese Kompetenz für den schulischen Bereich der Oberstufe und des Abiturs auch ohne den GeR in Deskriptoren zu fassen, war richtungsweisend. Die zahlreichen Aufgabenbeispiele, die dieser Fassung der Bildungsstandards beigefügt wurden, können Lehrkräften erläutern, was dieser Kompetenzbereich umfasst und welche Entwicklungspotenziale für sprachliche und persönlichkeitsbildende Aspekte ihm innewohnen. Dass es aber auch heute noch große Schwierigkeiten gibt zu bestimmen, wie valide und reliabel die Kompetenz der Lernenden dann beurteilt werden kann, ist Thema des Beitrags von Vogt.

Der abschließende Aufsatz von **Henning Rossa** (Dortmund) bezieht sich explizit auf das deutsche Bildungssystem: Wie verändern Bildungsstandards und zentrale Prüfungen den Fremdsprachenunterricht? Es handelt sich um eine Skizze eines Forschungsdesiderats zu intendierten und beobachteten Effekten der Standard- und Kompetenzorientierung. Rossa stellt seinen Beitrag also gewissermaßen unter das Motto: "The proof of the pudding is in the eating". Damit wird ein Brückenschlag vollzogen zwischen den allgemeineren testtheoretischen Beiträgen und der bildungspolitischen Perspektive der Schule.

Eva Burwitz-Melzer und Jürgen Quetz[2]

2 Korrespondenzadresse: Prof. Dr. Eva Burwitz-Melzer (Justus-Liebig-Universität Gießen, Eva.Burwitz-Melzer@anglistik.uni-giessen.de), Prof. i.R. Dr. Jürgen Quetz (Goethe-Universität Frankfurt am Main, jquetz@hotmail.de)