Mündliche Kompetenzen in der Fremdsprache fair messen

Überlegungen und Vorschläge zur Qualitätssicherung

Ulrike Arras | Bochum D

La producción e interacción oral en una lengua extranjera son actos sumamente complejos porque interactúan múltiples factores como conocimientos lingüísticos (vocabulario, gramática, pronunciación etc.), competencias pragmáticas y sociales y competencias interculturales. Además intervienen aspectos no verbales en la comunicación directa. Entonces, evaluar la comunicación oral en una lengua extranjera es un reto realmente difícil lo cual requiere de mucho cuidado. Este artículo tiene como objetivo sensibilizar en cuanto a algunos aspectos y criterios de calidad; tanto en la fase de diseño de exámenes orales (es decir el desarrollo de las tareas), como en la fase de la realización del mismo (por ejemplo los métodos de examinar, nuestra conducta como examinadora o examinador etc.). Y tambien sensibilzar en cuanto al uso adecuado de criterios de evaluación. El objetivo principal debería de ser el respeto de criterios importantes como la validez y la fiabilidad para así garantizar una cierta "fairness" como base para la evaluación de competencias lingüísticas.

„Mein Fall ist in Kürze dieser: Es ist mir völlig die Fähigkeit abhanden gekommen, über irgend etwas zusammenhängend zu denken oder zu sprechen."
(Hugo von Hofmannsthal*)

So weit sollten wir es nicht kommen lassen, schon gar nicht in einer mündlichen Prüfung! Wohl wahr: Sprechen, und insbesondere mündliche Interaktion, ist eine hochkomplexe Handlung, speziell in einer fremden Sprache. Diese Komplexität erklärt sich aus dem Zusammenspiel unterschiedlicher Faktoren, Kompetenzen und Wissensbestände. Da sind zum einen sprachsystemische Zugriffe auf Wortschatz, Sprachstruktur, Aussprache etc., weiters sind – kulturspezifisch geprägte – kommunikative und soziale Kompetenzen sowie Strategien der Gesprächsführung notwendig und nicht zuletzt interkulturelle Kompetenzen, etwa die sichere Verwendung des angemessenen Registers. Dazu kommen nonverbale Faktoren, die ebenfalls eine kulturspezifische Dimension besitzen.
Es ist daher eine besondere Herausforderung, mündliche Kompetenzen in der Fremdsprache zuverlässig, valide und somit fair zu messen. Die folgenden Ausführungen wollen für einige der Faktoren sensibilisieren, die dabei eine wichtige Rolle spielen und einen Beitrag liefern, um mündliche Sprachprüfungen einer Qualitätskontrolle zu unterziehen und ggfs. durch geeignete Massnahmen zu optimieren.

1. Prüfungsformen

Angesichts des Ziels, möglichst valide und authentisch mündliche Handlungskompetenz in der Fremdsprache – monologisches Sprechen und Interaktion zwischen zwei oder mehreren GesprächspartnerInnen – zu messen, werden meist kommunikative Performanztests in *face-to-face*-Prüfungssituationen gewählt. Diese Testform soll uns erlauben, anhand der gezeigten konkreten Äusserung (d.h. der Performanz) Hypothesen darüber zu entwickeln, wie ein Prüfling in ähnlich modellierten realen Situationen agieren würde. Daraus wird dann abgeschätzt, ob und in welchem Ausmass die Leistung den Anforderungen und Massstäben gemäss dem Testkonstrukt stand hält. Bei Performanztests stehen uns also direkt beobachtbare Leistungen zur Verfügung, während wir bei so genannten Kompetenztests nur indirekt auf die Fähigkeit der Kandidatin oder des Kandidaten schließen können. So können wir bei geschlossenen Itemtypen wie *Multiple-Choice*-Aufgaben zur Überprüfung des Leseverstehens in der Fremdsprache nur indirekt anhand der richtig oder falsch angekreuzten Lösung festzustellen versuchen, ob der Prüfling über die angestrebte Kompetenz verfügt. Man verwendet in diesem Zusammenhang auch oft die Dichotomie direktes versus indirektes Testen. Dank technischer Möglichkeiten hat sich in den letzten Jahren eine weitere Form etabliert, nämlich das semidirekte Testen. Bei dem sog. SOPI

(*Simulated Oral Proficiency Interview*) sind Aufgaben von einem Tonträger zu hören, auf den der Prüfling auch seine Antwort aufspricht. Die Kommunikation wird also simuliert und die Leistung als Konserve auf eine Kassette, eine CD oder auch auf Video gespeichert, um zu einem späteren Zeitpunkt und an einem anderen Ort abgehört und bewertet zu werden. Ein Beispiel für diesen Testtyp ist der Prüfungsteil „Mündlicher Ausdruck" der Prüfung TestDaF (Test Deutsch als Fremdsprache). Der Kritik, dass es sich dabei um keine echte Kommunikation handelt, weil alle Situationen und Sprechhandlungen simuliert sind, stehen jedoch gewichtige Vorteile gegenüber: Zum einen enthalten sie Aufgaben, die durch ihre situative Einbettung Merkmale authentischer Kommunikation zeigen, wenn Register und soziale Beziehung zwischen den fiktiven GesprächspartnerInnen vorgegeben sind. Ausserdem können verschiedene Sprechhandlungen elizitiert werden, sodass die Erfassung der Leistung nicht anhand einer einzigen Aufgabe erfolgt. Vor allem aber wird eine von situativen Faktoren und persönlichen Dispositionen der PrüferInnen unabhängige Beurteilung möglich. Beim TestDaF werden die weltweit aufgezeichneten Äusserungen zentral in Deutschland von eigens dafür geschulten BeurteilerInnen bewertet. Zudem werden alle Prüflinge zum selben Zeitpunkt mit denselben Aufgaben konfrontiert – ein für die Fairness eines High-Stakes-Tests immens wichtiger Punkt. Ein dritter Vorteil liegt in der Standardisierung, denn die Zahl und Art der geforderten Sprechhandlungen, ihre situativen Faktoren, die fiktiven GesprächspartnerInnen etc. sind stets gleich, sodass die Schwierigkeit bei jedem Testereignis annähernd konstant ist. So sollen alle Prüflinge bei Aufgabe 2 im Gespräch mit einer deutschen Kommilitonin oder einem deutschen Kommilitonen stets ein eigenkulturelles Phänomen beschreiben. Auch die zur Verfügung stehende Zeit ist genau festgelegt und auf die jeweilige Aufgabe ausgerichtet. Auf diese Weise lassen sich die Schwierigkeitsdeterminanten sehr genau kontrollieren, zumal auch die Beurteilungsmassstäbe testsatzübergreifend konzipiert sind.

Am meisten verbreitet sind freilich direkte Testformen, Einzel- oder Paarprüfungen, manchmal auch Gruppenprüfungen in *face-to-face*-Situationen. Dieses Vorgehen erlaubt zum einen eine Rollenverteilung seitens der PrüferInnen, bei der eine Person vorbereitete Fragen stellt und das Gespräch führt, während die andere eine beobachtende Rolle einnimmt, was eine gewisse Objektivität gewährleistet. In der traditionellen Konstellation – PrüferIn als GesprächspartnerIn und BewerterIn in Personalunion zusammen mit einem Prüfling – sind diese parallelen Tätigkeiten so komplex und von so vielen unerwünschten Faktoren geprägt, dass weder eine faire Durchführung der Prüfung noch eine zuverlässige Evaluation möglich ist.

Ein Vorteil der Paarprüfung ist, dass eine weitgehend symmetrische Konstellation zwischen den beiden Prüflingen fingiert werden kann, so dass die kommunikative Situation authentischer wirkt (jedoch keineswegs unbedingt ist!). Auf diese Weise kann man feststellen, inwiefern bestimmte kommunikative Strategien und pragmatische Kompetenzen beherrscht werden, etwa angemessenes turn-taking, die Eroberung des Rederechts oder auch Sprechhandlungen wie Argumentieren oder Fragenstellen. Ähnliches kann im Rahmen einer Gruppenprüfung erzielt werden, jedoch mit dem Nachteil, dass die Äusserungen und Leistungen der einen Person jene der anderen beeinflussen. Das kann sich positiv auswirken, wenn ein echtes Gespräch entsteht, das Team gut eingespielt ist und Empathie zwischen den Prüflingen vorherrscht, aber auch negativ, wenn Konkurrenzdenken die Situation beherrscht und wenn die Prüflinge nicht miteinander, sondern zur Prüferin oder zum Prüfer hin orientiert sprechen, um vielleicht eine bessere Einstufung zu erhalten. Ungünstige Auswirkungen entstehen auch bei einer grösseren Kompetenzdiskrepanz zwischen beiden Prüflingen.

Aus Kapazitätsmangel (zu wenige PrüferInnen auf der einen, zu viele zu testende LernerInnen auf der anderen Seite, zudem Zeit- oder Raummangel) entscheiden Institutionen häufig, mehrere Prüflinge zusammen in einer Gruppe *face-to-face* zu testen. Auch hier stehen einigen Vorteilen gewichtige Nachteile gegenüber. Einerseits können interaktive kommunikative Sprechhandlungen inszeniert werden, z.B. Rollenspiele, andererseits ist es nahezu unmöglich, die Schwierigkeit der Sprechaufgaben für alle gleich zu gestalten. Prinzipiell ist bei Paar- und Gruppenprüfungen besonderes Augenmerk darauf zu legen, dass alle die gleichen Chancen bekommen, um ihre Kompetenzen unter Beweis zu stellen.

2. Aufgabentyp

Üblicherweise werden mehr oder minder offene Aufgabentypen eingesetzt, wenn es gilt, mündliche Kompetenzen in der Fremdsprache zu messen. Hierbei stellen i. d. R. eine Prüferin oder ein Prüfer Fragen in der Zielsprache, die auch in dieser zu beantworten sind. Zentrales Merkmal ist der Grad der Steuerung: Wie viel Eigenaktivität wird den KandidatInnen zugestanden? Ist das Ganze ein Frage-Antwort-Spiel, bei dem in oftmals rascher Form mit klar verteilten Rollen agiert wird (was dann nicht selten wie ein Verhör wirkt)? Oder handelt es sich um eine Diskussion oder ein (Fach-) Gespräch, bei dem der Prüfling weit mehr von sich selbst und seinen Überlegungen einbringen und u. U. sogar selbst eine Frage einwerfen kann?

3. Schwierigkeitsfaktoren

Ein weiteres Problem liegt in der Schwierigkeit der Aufgabenstellung, die sich zum einen aus deren Determinanten ergibt, also beispielsweise aus Faktoren wie:

- Thema – abhängig davon, wie viel thematisches Vor-, Welt- oder auch Fachwissen für die Umsetzung der Aufgabe erforderlich ist,
- Register – je nach der fiktiven kommunikativen Situation, in die die Aufgabe eingebettet ist,
- geforderte Sprechhandlungen – kognitiv anspruchsvoll wie z.B. Argumentieren (CALP) versus informelles Erzählen (für das eher BICS benötigt werden).[11]

Dabei können wir nicht einfach einen Faktor herausgreifen und daran feststellen, ob eine Aufgabe schwierig oder leicht ist. Denn die zur Verfügung stehende Zeit zusammen mit anderen Faktoren können eine Aufgabe mehr oder weniger schwierig machen. Prinzipiell gilt: „Difficulty is not a direct characteristic of tasks; rather it is the sum of task characteristics and the conditions under which someone performs the task (standing on their head or on their two feet) in relation to the person's ability in the skills that it requires." (Luoma, 2004: 46). Zudem spielt die Beurteilung eine Rolle: Eine eher leichte Aufgabe kann durch eine eher strikte Beurteilung ein höheres Schwierigkeitsniveau erreichen, in der umgekehrten Konstellation kann eine schwierige Aufgabe durch eine mildere Beurteilung ausgeglichen werden.

4. Beurteilung

Ausgangspunkt eines jeden Tests ist das Testkonstrukt, also jene Aspekte von Kompetenz, die gemäss der Funktion, der Zielgruppe und der Zielsetzung der Prüfung vorgegeben sind; diese müssen sich in der Aufgabenstellung sowie in den Beurteilungskriterien widerspiegeln. Es ist evident, dass ein Test zur Messung kommunikativer mündlicher Kompetenz auf B2 des Gemeinsamen Europäischen Referenzrahmens auch tatsächlich die diesem Niveau entsprechenden Aspekte überprüfen muss. Ebenso kann ein Test, der die Beherrschung grammatischer Strukturen mittels halboffener Items (Lückentext) überprüfen soll, natürlich nicht gleichzeitig kommunikative mündliche Kompetenzen messen. Auch wenn die Beurteilungsmassstäbe in erster Linie auf grammatische Fehler oder Akzentfreiheit ausgerichtet sind, ist kommunikative Kompetenz nicht valide zu erfassen.

Ein zentrales Anliegen bei der Beurteilung ist die Zuverlässigkeit; sie sollte möglichst reliabel, also von der beurteilenden Person unabhängig sein. Wird ein und dieselbe Leistung von unterschiedlichen Personen oder von derselben Person zu unterschiedlichen Zeitpunkten unterschiedlich bewertet, dann liegt eine schwache Inter-Rater- bzw. Intra-Rater-Reliabilität vor. In beiden Fällen spielen offensichtlich Faktoren mit, die für die Messung mündlicher Kompetenzen irrelevant sind, z.B. Vorlieben oder Antipathien, Einstellungen oder das akute Befinden der beurteilenden Person. Ebenso können der Akzent eines Prüflings, seine Stimme, Gestik oder Mimik, aber auch inhaltliche Aspekte seiner Äusserung die Wahrnehmung und somit die Bewertung beeinflussen.[2]

Der Persönlichkeitsfaktor, der sicher am stärksten ins Gewicht fällt, ist die individuelle Strenge oder Milde. Seit einigen Jahren steht uns nun ein Instrument zur Verfügung, das Multifacetten-Rasch-Modell von Linacre (2004), das von Testinstitutionen wie dem TestDaF-Institut systematisch zum Ausgleich verschiedener Faktoren auf die Beurteilung eingesetzt wird. Hierbei wird ein individueller Strengekoeffizient ermittelt und bei der endgültigen Einstufung der Leistung berücksichtigt, wodurch man den sog. fairen Durchschnitt erreicht. Dieser gibt für eine Person „ihre um die Strenge bzw. Milde der involvierten Beurteiler wie auch um die Schwierigkeit des jeweiligen Kriteriums bzw. der jeweiligen Aufgabe bereinigte mittlere Einstufung in der Metrik der Ratingskala an." (Eckes, 2003: 59)

Doch selbst ohne solche Instrumente lassen sich Verzerrungen in der Bewertung kontrollieren. Rückgrat jeder Beurteilung sollten die Kriterien sein, die unbedingt vorab zu entwickeln sind und die Anforderungen der gestellten Aufgaben und das Testkonstrukt widerspiegeln sollten. Es ist absolut unzulässig, Beurteilungskriterien während oder nach einer mündlichen Prüfung mehr oder minder ad hoc oder auf individuellen Erfahrungen basierend zu entwickeln, denn dann sind die Massstäbe tendenziell individuell konzipiert, werden nicht unabhängig von dem Prüfling angelegt oder werden interpersonell unterschiedlich angewendet.

Praktische Orientierung bei der Entwicklung und Zusammenstellung geeigneter Kriterien bieten Referenzsysteme wie der schon genannte Gemeinsame Europäische Referenzrahmen für Sprachen (GER). Er enthält eine Vielzahl von meist empirisch abgesicherten Skalen (sprachliches Spektrum, Register, Interaktion, monologisches Sprechen u.v.m.), die man je nach Bedarf übernehmen oder den spezifischen Anforderungen, Lerninhalten und Zielen des Kurses anpassen kann.[3] Diese verhelfen zu einem möglichst hohen Mass an Reliabilität, denn die Leistungen werden stets nach denselben Kriterien beurteilt, unabhängig von Prüfungsereignis und -termin, von den Aufgaben und Fragen der aktuellen Prüfung sowie von den PrüferInnen. Ein Training ist allerdings unbedingt erforderlich, um sich über die Anwendung der Skalen zu verständigen. Zwar wird auch die Zuhilfenahme dieser Hilfsmittel keine umfassende Inter-Rater-Übereinstimmung gewährleisten, dennoch führt die Handhabung in der *peer group* – im Kollegium bzw. zusammen mit den an der mündlichen Prüfung beteiligten PrüferInnen – zu mehr Sicherheit und einem gewissen Konsens. Äusserst nütz-

Grundsätzlich beeinflussen die eigenen beruflichen, fremdsprachlichen und lebensweltlichen Erfahrungen unsere Wahrnehmung und damit unser Urteil über fremdsprachliche Leistungen.

lich sind dabei die sog. *benchmarks*, also die konkreten Beispiele, die mit den jeweiligen Beurteilungen auch die Anwendung der GER-Skalen veranschaulichen.

Wenn wir gesprochene Sprache valide testen wollen, dann müssen wir auch deren spezifische Merkmale, insbesondere im Dialog, angemessen berücksichtigen. Diese können wir nicht an schriftsprachlichen Kategorien der Korrektheit ausrichten, sondern müssen gerade charakteristische Aspekte wie Satzabbrüche, Selbstkorrekturen, Verzögerungen, Topikalisierungen und auch grammatikalische Abweichungen von der Schriftsprache annehmen. Zudem ist zu klären, ob und wie phonetische Besonderheiten wie Aussprache, Intonation, Satz- und Wortakzent, aber auch der Redefluss bewertet werden sollten. Was den durch die Erstsprache geprägten Akzent anbelangt, sei wiederum auf den GER verwiesen, der eine – freilich recht grob strukturierte – Skala zum Bereich der phonologischen Kompetenz liefert. Die Leistungsbeschreibungen sehen dort Akzent bis zum Niveau B1 als akzeptabel an. Ausschlaggebend ist – gemäss der Kompetenzorientierung des GER – die Verständigung: Im Mittelpunkt steht die Kommunikation, die trotz eines Akzents oder bestimmter Intonationsfehler durchaus gewährleistet sein kann. Auch in den Beispielen der dem GER beigefügten *benchmarks* haben alle SprecherInnen einen teils starken, muttersprachlich geprägten Akzent. Das sollte jedoch nicht zu Abwertungen bei der Leistungsbeurteilung führen – es sei denn, es handelte sich um Aussprachefehler wie einen sinnentstellenden Wortakzent, der leicht zu Missverständnissen führen kann, etwa im Falle von Homografen wie módern versus modérn.
Freilich ist gerade Akzent ein Faktor, der auf Rezeptionsseite individuell als problematisch oder aber als angenehm wahrgenommen und beurteilt wird. So werden bekanntermassen bestimmte Akzente durchaus als attraktiv empfunden, was wohl in erster Linie mit dem Prestige der Ausgangssprache zu tun hat wie z.B. der französische Akzent in Deutschland. Andererseits kann dadurch auch die Kommunikation stark beeinträchtigt werden bis hin zu Konzentrationsmangel und Verweigerung durch den Hörer. Für Lehrkräfte zu beachten ist eine besondere Art der *déformation professionnelle*, bei der sie sich langsam an den Akzent von SprecherInnen bestimmter Zielkulturen gewöhnen, vor allem, wenn sie selbst die Ausgangssprache der Lernenden beherrschen oder sogar in dem jeweiligen Kultur- und Sprachraum gelebt haben.

Grundsätzlich beeinflussen die eigenen beruflichen, fremdsprachlichen und lebensweltlichen Erfahrungen unsere Wahrnehmung und damit unser Urteil über fremdsprachliche Leistungen. Aus diesem Grunde könnte man sich zu Recht fragen, ob es eigentlich eindeutige Massstäbe überhaupt geben kann.

5. *Que faire?* Massnahmen zur Qualitätssicherung

An erster Stelle steht sicher das Schulen und das Monitoring der BeurteilerInnen bzw. PrüferInnen. Prinzipiell ist es ratsam, die *peer group* (Arbeitsgruppe, Kollegium, Prüfungsteam) einzubinden, denn der Austausch untereinander führt zu grösserer Sensibilisierung für individuell entwickelte Beurteilungsstrategien und so zu deren besseren Kontrolle.
Eine zweite dringend empfehlenswerte Massnahme ist die Erstellung einer Beispielsammlung wie die schon erwähnten *benchmarks,* am besten per Video, für die verschiedenen Leistungsniveaus. Flankierend sollten gegenseitige Hospitationen in mündlichen Prüfungen erfolgen. Hilfreich für die Kalibrierung ist es zudem, wenn die entsprechenden Arbeitsgruppen heterogen zusammengesetzt sind und auch – sozusagen als Korrektiv – KollegInnen aus anderen Fachdidaktiken umfassen. So könnte etwa das o. g. Problem einer individuell geprägten Wahrnehmung von Akzent durch eine Gruppe, in der Fremdsprachen-Lehrkräfte und „normale" HörerInnen vertreten sind, relativiert werden.
Zukunftsweisend ist es in jedem Fall, sich sprachenübergreifend bzw. interinstitutionell über Massstäbe und *benchmarks* zu verständigen. Nur auf diese Weise kann geklärt werden, inwiefern eine B1-Leistung im Italienischen jener B1-Leistung im Deutschen oder Chinesischen entspricht. Aus derartigen Überlegungen und Verständigungen erwachsen sodann curriculare Konsequenzen, die am Ende in ein vereinheitlichtes Anforderungsprofil münden sollten, um schliesslich Charles Aldersons provokative Frage „Is your B1 my B1?" positiv beantworten zu können.
Von zentraler Bedeutung sind ohne Zweifel die Bedingungen, unter denen mündliche Prüfungen durchgeführt werden. Insbesondere sollte auf eine angenehme Atmosphäre geachtet werden, die den Prüflingen erlaubt, wirklich das zu zeigen, was sie können. Daraus ergibt sich das Prinzip einer positiven Herangehensweise: Ziel der Prüfung sollte sein, die Kompetenzen der KandidatInnen zu ermitteln, nicht aber herauszufinden, was sie nicht können. Fehlersuche, verhörähnliche Methoden, Fangfragen und Fallen sind deshalb fehl am Platze.

Mündliche Prüfungen erfordern somit geeignete Frage- und Gesprächstechniken. Schwierig ist es dabei zu unterscheiden, inwiefern wir lediglich sprachliche Kompetenzen oder nicht auch inhaltliches Wissen oder Einstellungen messen. Je höher das Sprachniveau und je komplexer die Inhalte, desto stärker interagieren beide Bereiche. Sprache und Inhalt sind nicht zu trennen, eine Trennung dient lediglich der Systematisierung.

Ebenso muss man sich bewusst sein, dass nonverbale Faktoren einen erheblichen Teil der direkten Kommunikation steuern und damit Verstehen sichern oder erschweren, vor allem bei interkulturellen Unterschieden und einer asymmetrischen Machtverteilung im Gespräch (Gefälle zwischen Prüferin bzw. Prüfer und Prüfling). Zeigt der Prüfer durch Mimik und Körperhaltung Desinteresse und Mangel an Aufmerksamkeit, so beeinträchtigt dies selbstredend die Konzentration und die Performanz des Kandidaten. Umgekehrt können durch Nervosität und Unsicherheit hervorgerufene Gesten, Körperhaltungen oder Stimmlagen des Prüflings die Rezeption der Prüferin (negativ) beeinflussen. Alle diese Faktoren verzerren sowohl die Leistung als auch die Beurteilung. Sie sind zwar kaum gänzlich zu vermeiden, jedoch können auch hier Monitoring und Schulungen für Sensibilisierung sorgen und das Problem zumindest relativieren. Nützlich zur Aktivierung und Entwicklung selbstreflexiver Fähigkeiten sind Laut-Denken-Verfahren (Arras, 2007a und Arras et al., 2009). Generell sollten alle Äusserungen der PrüferInnen, seien sie verbal oder nonverbal, positiv sein: Wir müssen Geduld aufbringen, unterstützen, uns selbstverständlich ironische Bemerkungen usw. verkneifen, ebenso alle irrelevanten Aktivitäten wie Papierrascheln, Notizen machen, Unterlagen suchen usw. vermeiden. Auch dürfen wir selbst nicht zu viel sprechen oder sogar unsere Fragen selbst beantworten. Ein besonderes Problem ist die Fehlerkorrektur: Eine Prüfung hat nicht unbedingt Lernzuwachs zum Ziel, und die Korrektur sprachlicher Fehler bleibt in einer Situation, wenn der Kandidat nervös und weder emotional noch motivational lernbereit ist, sicherlich ohne Lerneffekt. Unsere Korrekturbemühungen sind dann im besten Fall zwecklos, oft verunsichern sie zusätzlich und werden als Negativ-Feedback wahrgenommen.

Zusammenfassend lässt sich festhalten: Obwohl gerade das Messen mündlicher Kompetenzen in der Fremdsprache aufgrund der vielen interagierenden Faktoren eine besondere Herausforderung darstellt, lässt sich durch die Berücksichtigung einiger entscheidender Aspekte und durch Massnahmen zur Qualitätssicherung sowohl die Testkonzeption als auch die Durchführung und damit die Beurteilung der Leistungen optimieren, um zentrale Testgütekriterien und schliesslich Fairness zu gewährleisten. Grundsätzlich sollten wir alle Phasen so gestalten, dass sie eine objektive, genaue und reliable Erfassung der tatsächlichen Kompetenzen ermöglichen. Dies verlangt freilich einiges an Aufwand und Mühe seitens der Testerstellung, der Testdurchführung und der Leistungsbeurteilung. Aber sich Mühe geben, das verlangen wir ja auch von unseren Prüflingen!

Anmerkungen

* *Ein Brief*, zitiert nach: *Gesammelte Werke in zehn Einzelbänden, Bd. 7: Erzählungen, erfundene Gespräche und Briefe, Reisen*. Frankfurt a. M.: Fischer 1979, S. 465).

[1] Zu dem aus der Bilingualismus-Forschung stammenden Begriffspaar BICS (*basic interpersonal communication skills*), also ein allgemeinsprachlicher Sprachgebrauch, der von *context-embedded communication* gekennzeichnet ist, versus CALP (*cognitive academic language proficiency*), das bedeutet ein im akademischen Kontext erforderlicher Sprachgebrauch, der hohen kognitiven Anforderungen Genüge leisten muss, u. a. weil er *context-reduced communication* eignet, s. Cummins/Swain 1986.

[2] Vgl. in diesem Zusammenhang den Beitrag von Kerstin Reinke in diesem Heft.

[3] Wie eine solche Adaptation aussehen kann, zeigen Hannele Kara in ihrem Beitrag über mündliche Leistungsbewertung an finnischen Schulen und Gregor Chudoba in seiner Bewertungsskala von der Universität Klagenfurt in dieser Nummer.

La "Bocca della verità" a Roma.

Literatur

Althaus, H.-J. (2004). Der TestDaF. In DAAD (ed.). *Die internationale Hochschule: Ein Handbuch für Politik und Praxis.* Bd. 8, 80–87. Bielefeld: Bertelsmann.

Arras, U. (2009). Kompetenzorientierung im Fremdsprachenunterricht – was heißt das eigentlich? *Pandaemonium Germanicum. Revista de estudos germanísticos* 14/2009, 206–217. Universidade São Paulo.

Arras, U. (2007a). *Wie beurteilen wir Leistung in der Fremdsprache? Strategien und Prozesse bei der Beurteilung schriftlicher Prüfungsleistungen am Beispiel des TestDaF (Test Deutsch als Fremdsprache).* Giessener Beiträge zur Fremdsprachenforschung. Tübingen: Narr.

Arras, U. (2007b). Zur Revision des Subtests ‚Mündlicher Ausdruck' der Prüfung ‚Test Deutsch als Fremdsprache' (TestDaF). *Studienkolleg. Zeitschrift zur Pädagogik und Didaktik studienvorbereitender Kurse für ausländische Studierende*, 11/2007, 5–29.

Arras, U. (2006a). Was macht eine Aufgabe schwierig? Schwierigkeitsdeterminanten in Tests zur Überprüfung mündlicher Kompetenzen am Beispiel von TestDaF-Aufgaben des Subtests Mündlicher Ausdruck. *Lingua Franca – Lingua Academica, Mehrsprachigkeit im Europäischen Hochschulraum*, 211–225. Bochum: AKS.

Arras, U. (2006b). Der TestDaF. Konzept und Prinzipien des standardisierten Tests Deutsch als Fremdsprache. *Fòrum – Anuari de l'Associació de Germanistes de Catalunya. Akten des sechsten Kongresses des Katalanischen Deutschlehrer- und Germanistenverbandes (A.G.C.)*, 39–52. Tarragona.

Arras, U., Marks, D. & Zimmermann, S. (2009). BeurteilerInnen in den Kopf geschaut: Wie das Verfahren des Lauten Denkens im Rahmen von Beurteilungsschulungen eingesetzt werden kann. *DaF-Brücke: Zeitschrift für Deutschlehrerinnen und Deutschlehrer Lateinamerikas*, 11/2009: 5–9.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing* 19 (4), 453–476.

Bachman, L. F. & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bolton, S., Glaboniat, M., Lorenz, H., Müller, M., Perlmann-Balme, M. & Steiner, S. (2008). *Mündlich. Mündliche Produktion und Interaktion Deutsch. Illustration der Niveaustufen des Gemeinsamen Europäischen Referenzrahmens.* Berlin/München: Langenscheidt.

Cummins, J. & Swain, M. (1986). *Bilingualism in education. Aspects of theory, research and practice*. London/New York.

Eckes, T. (2010). Die Beurteilung sprachlicher Kompetenz auf dem Prüfstand: Fairness in der beurteilergestützten Leistungsmessung. In K. Aguado, K. Schramm & H. J. Vollmer (eds.), *Fremdsprachliches Handeln beobachten, messen, evaluieren: Neue methodische Ansätze der Kompetenzforschung und der Videographie* (pp. 65–97). Frankfurt: Lang.

Eckes, T. (2005). Evaluation von Beurteilungen: Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell. *Zeitschrift für Psychologie,* 213, 77–96.

Eckes, T. (2004). Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In Wolff, A., Ostermann, T. & Chlosta, C. (eds.), *Integration durch Sprache (Materialien Deutsch als Fremdsprache*, Bd. 73, pp. 485–518). Regensburg: Fachverband Deutsch als Fremdsprache.

Eckes, T. (2003). Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse. *Fremdsprachen und Hochschule*, 69, 43–68.

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

Fulcher, G. (2003). Testing Second Language Speaking. London et al.: Pearson/Longman.

Grotjahn, R. & Tesch, B. (2010). Messung der fremdsprachlichen Sprechkompetenz im Fach Französisch. In Porsch, R., Tesch, B. & Köller, O. (eds.) (2010). *Standardbasierte Testentwicklung und Leistungsmessung* (pp. 177–205). Münster et al.: Waxmann.

Kenyon, D. M. (2000). Tape-mediated Oral Proficiency Testing: Considerations in Developing Simulated Oral Proficiency Interviews (SOPIs). In Bolton, S. (ed.) (2000). *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*, (pp. 87–106). München: Goethe-Institut.

Kniffka, G. & Üstünsöz-Beurer, D. (2001). TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassettengesteuerten Testformats. *Fremdsprachen Lehren und Lernen*, 30, 127–149.

Krause, W. D. & Sändig, U. (2002). *Testen und Bewerten kommunikativer Leistungen im Unterricht Deutsch als Fremdsprache: Linguistische Grundlagen und didaktische Angebote*. Frankfurt/Main: Lang.

Lazarton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests. (= Studies in Language Testing 14)*. Cambridge: Cambridge University Press.

Linacre, J. M. (2004). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: MESA press.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: CUP.

Österreichisches Sprachdiplom Deutsch (ed.) (1999). *Österreichisches Sprachdiplom Deutsch – Zertifikat Deutsch: Modellsatz*. Wien.

Tschirner, E. (2005). Das ACTFL OPI und der Europäische Referenzrahmen. *Babylonia* 2, 50–55.

Tschirner, E. (2001a). Die Evaluation fremdsprachlicher mündlicher Handlungskompetenz. *Fremdsprachen Lehren und Lernen* 30, 87–115.

Tschirner, E. (2001b). Die ACTFL Leitlinien mündlicher Handlungsfähigkeit. *Fremdsprachen Lehren und Lernen* 30, 116–126.

Internetseiten

TestDaF-Institut: www.testdaf.de

Association of Language Testers in Europe (ALTE): www.alte.org

Ulrike Arras

ist Referentin für Testentwicklung am TestDaF-Institut, das seit 2001 die standardisierte Prüfung Test Deutsch als Fremdsprache erstellt und weltweit administriert. Ihr Arbeitsgebiet umfasst neben Fragen der Testerstellung und Leistungsbeurteilung die Durchführung von Schulungen und Fortbildungen. Sie war außer in Deutschland auch länger in der VR China, in Spanien, Marokko, Ägypten und Venezuela tätig. Sie hat Sprachlehrforschung, Germanistik, Sinologie sowie Politische Wissenschaften Südostasiens studiert und im Fach Sprachlehrforschung über spezifische Strategien bei der Beurteilung schriftlicher Prüfungsleistungen promoviert.