Einleitung

Das vorliegende Studienbuch ist an Lehramtsstudierende in Ausbildung sowie an ReferendarInnen, Schul- und UnterrichtspraktikantInnen in Deutschland und Österreich, der Schweiz und Südtirol gerichtet, die an einer Universität und einer Hochschule Fremdsprachen oder die klassischen Sprachen Latein oder Griechisch studieren. Entstanden ist das Studienbuch an der Universität Innsbruck. Hier werden am sog. Innsbrucker Modell der Fremdsprachendidaktik (IMoF) künftige FremdsprachenlehrerInnen seit dem Studienjahr 2001 / 2002 sprachenübergreifend und sprachspezifisch ausgebildet. IMoF widmet sich schulischer Mehrsprachigkeit und multilingualen Herangehensweisen in schulischen Kontexten und wird über Innsbruck und Österreich hinaus als Meilenstein einer sprachenintegrierenden fachdidaktischen Ausbildung gewürdigt (Krumm & Reich 2013; s. auch BMUKK & BMWF 2008, 48 f.), die den Rahmen monolingualer Studiengänge hinter sich lässt und sprachenverbindende sowie mehrsprachigkeitsdidaktische Ansätze in den Fokus rückt.

Das Modell hat seine Anfänge im Jahr 2000, als ein neuer Studienplan für die Ausbildung künftiger FremdsprachenlehrerInnen an der Universität Innsbruck zu konzipieren war. Es stellte sich für den fremdsprachlichen Unterricht die Frage, ob es Theorien, Grundlagen und Prinzipien gibt, die jeweils nur auf eine Zielsprache zutreffen, oder ob nicht vielmehr Theorien, Grundlagen und Prinzipien der sprachdidaktischen Vermittlung allen Zielsprachen gemein sind. Auf Initiative von Barbara Hinger kamen FremdsprachendidaktikerInnen der Institute für Anglistik, Romanistik und Slawistik bei Diskussionen am Runden Tisch schließlich überein, dass Synergien nicht von der Hand zu weisen sind: Gemeinsame sprachenübergreifende Theorien und Grundlagen des Unterrichtens moderner Sprachen sind deutlich auszumachen, diese wären aber auch auf einzelsprachliche Inhalte zu spezifizieren, um den Unterricht in einer konkreten Zielsprache adäquat umsetzen zu können. Diese zweifache Perspektive, gebündelt in der Kombination von sprachenübergreifender und sprachspezifischer Fremdsprachendidaktik, sollte sowohl das Erarbeiten theoretischer Ansätze und empirischer Forschungsergebnisse als auch deren reflektierte Übertragung in den schulischen Alltag gewährleisten.

Dieselbe Herangehensweise wurde auf Anregung von Wolfgang Stadler auf den Bereich des Prüfens und Bewertens von Fremdsprachen übertragen und in das Curriculum integriert: Bis dahin war dieses Gebiet in der Ausbildung kaum vorgesehen, obwohl Lehrpersonen ihr gesamtes Berufsleben hindurch die sprachlichen Leistungen von SchülerInnen zu bewerten haben (vgl. Arras 2009, 169, die von der Beurteilung fremdsprachlicher Leistungen als dem „täglichen Brot" aller Lehrkräfte spricht), Klassen- / Schularbeiten, Tests und mündliche Prüfungen erstellen, deren Ergebnisse auswerten und auf der Basis dieser sowie formativ bewerteter Leistungen zu einer summativen Gesamtbeurteilung für jede / jeden SchülerIn am Ende eines Lernjahres gelangen müssen. Die dafür nötigen Kompetenzen (*assessment literacy*) sollten in entsprechenden Lehrveranstaltungen erworben werden. Diese Argumente führten dazu, auch den Bereich des Testens und Bewertens fremdsprachlicher Kompetenzen in das Konzept der neuen Lehramtsausbildung aufzunehmen und eine sprachenübergreifende, theoriebasierte Lehrveranstaltung „Testen und Bewerten" zu konzipieren, die von sprachspezifischen und schulbezogenen Begleitkursen flankiert wird.

2002 wurde die erste sprachenübergreifende „Einführung in die Didaktik des Fremdsprachenunterrichts" im *team teaching*-Verfahren umgesetzt; für die sprachspezifischen Begleitkurse konnten schulische Lehrpersonen gewonnen werden, die ihre praktische Erfahrung einfließen ließen und sich durch die Kooperation mit Lehrenden an der Universität auch weiter professionalisieren konnten. Darüber hinaus wurde IMoF Motor für eine zuvor nur marginal existierende Forschung in der Fremdsprachendidaktik an der Universität Innsbruck[1]. Bereits 2002, im ersten Semester der Durchführung, wurde das Modell mit dem „Europasiegel für innovative Sprachenprojekte" ausgezeichnet.

2015 wurde – im Zuge der Neukonzipierung der Curricula als Bachelorstudiengänge – die Präsenzzeit für die Lehrveranstaltung „Einführung in das Testen und Bewerten von Fremdsprachen" erhöht. Diese Erweiterung basiert in nicht unwesentlichem Ausmaß auf dem Feedback von Studierenden, die die Bedeutung dieser Thematik für ihr späteres Berufsfeld erkannten und in Befragungen entsprechend hervorhoben. In der Dissertation von Hirzinger-Unterrainer (2013), die IMoF aus Sicht der Studierenden evaluierte, konnte für das Abschlussmodul „Testen und Bewerten" Folgendes festgehalten werden:

> Das ganze Abschlussmodul erachtet [eine Studierende] als sehr wichtig, sie habe sich „[…] *nämlich nie die Frage gestellt, wie stelle ich einen Test zusammen*" … Die Lehrveranstaltung, aber vor allem das [begleitende] Korrekturpraktikum, habe sie zum Nachdenken über geeignetes Testen und Bewerten angeregt. Das Wissen aus diesem Modul erachte sie für ihren späteren Beruf als sehr bedeutend. (ebd., 293)

Dass adäquates Heranführen an Prinzipien des Testens und Bewertens fremdsprachlicher Leistungen grundsätzlich von Studierenden geschätzt wird und sie diesem Bereich in ihrer Ausbildung großen Wert beimessen, zeigt folgendes Zitat:

> [Studierende geben] den Wunsch an, durch dieses Modul gegen Ende des Studiums Sicherheit in der Notengebung zu erlangen. […] Da die Studierenden eine große Unsicherheit im Bereich Testen und Bewerten spüren, sind sie für die vermittelten Hilfestellungen dankbar. (ebd., 356)

In den Augen der beteiligten FremdsprachendidaktikerInnen hat die Beschäftigung mit dem Testen und Bewerten fremdsprachlicher Kompetenzen auch ihre eigene Professionalisierung vorangetrieben und das Teambewusstsein gestärkt: So absolvierten die Verantwortlichen der sprachspezifischen Begleitworkshops gemeinsam eine Fortbildung im kommunikativen Sprachentesten an der *Lancaster University* in England. Damit entstand neben einer positiven Gruppendynamik im Erwerb und der Erweiterung ihrer Expertise auch eine Vertiefung ihrer Sprachbewertungskompetenz (*language assessment literacy*), die mittlerweile international in unterschiedlichsten Kontexten gefordert wird (vgl. u. a. Harsch 2015, Harding & Kremmel 2016). Einige Teammitglieder sowie junge IMoF-AbsolventInnen erwarben einen ebenfalls von der *Lancaster University* angebotenen Online-Master in *Language Testing*, andere haben

1 Hintergründe, theoretische Basierungen sowie empirische Einblicke zu IMoF gewähren Publikationen wie Hinger (2009a, 2016a), Hinger & Schmiderer (im Druck) oder Hirzinger-Unterrainer (2013, 2014a); s. auch https://tinyurl.com/y9s3z2ml (21. 09. 2017).

an Ausbildungen in *Item Writer Training*-Seminaren teilgenommen und sind ExpertInnen für die Erstellung kriterienorientierter Aufgabenformate im Rahmen der mittlerweile flächendeckend an österreichischen Schulen der Sekundarstufe II eingeführten und gesetzlich verankerten standardisierten, teilzentralen und kompetenzorientierten Reife- und Diplomprüfung (SRDP) in den Fremdsprachenfächern geworden. Carol Spöttl, die zu Beginn die einzige Expertin im Sprachentesten an der Universität Innsbruck war, etablierte eine *Language Testing Research Group Innsbruck* (LTRGI[2]), im Rahmen derer Forschungsprojekte lukriert, junge AbsolventInnen in der Sprachtestforschung verankert und nationale wie internationale Vernetzungen geschaffen werden konnten: Erwähnt sei an dieser Stelle die Organisation der 9. Tagung von EALTA-*European Association of Language Testing and Assessment* 2012 und die Umsetzung der 4. *Summer School* von EALTA 2016.

Das vorliegende Buch spiegelt zu einem großen Teil Inhalte des IMoF-Moduls „Testen und Bewerten" wider, geht aber in einigen Kapiteln darüber hinaus. Ausbildungsinhalte beziehen sich auf unterschiedliche Funktionen sprachlicher Leistungsbeurteilung und ihre gesetzlichen Vorgaben im schulischen Kontext, auf die für das Überprüfen von Sprachen wesentlichen Testgütekriterien, auf Konstruktdefinitionen für sprachliche Fertigkeiten und sprachliche Mittel oder auf kontinuierliches Bewerten sprachlicher Leistungen. Die Lehrveranstaltung wird im sprachenübergreifenden Team geplant und teilweise gemeinsam, teilweise individuell umgesetzt. Konkret bedeutet dies, dass Studierende das Erstellen adäquater Aufgabenformate für die unterschiedlichen sprachlichen Fertigkeiten und sprachlichen Mittel, bezogen auf verschiedene Sprachniveaus, ebenso erlernen wie das Erstellen von Klassen- und Schularbeiten für bestimmte Lernjahre. Indem sie verschiedene Bewertungsarten und -raster kritisch reflektieren und gemeinsam diskutieren, wird der für das Verfassen von Prüfungsaufgaben wichtige kooperative Charakter betont und für Studierende bereits im Studium konkret erfahrbar. Die spätere Zusammenarbeit von Fremdsprachenlehrpersonen an der Schule soll so im Studium präjudiziert und erlernt werden. Ob dies durch die IMoF-Ausbildung in der späteren Unterrichtspraxis der AbsolventInnen auch gelingt, können nur entsprechende Langzeitstudien zeigen. Jenseits von IMoF bleibt die Ausbildung für schulische Mehrsprachigkeit sowie für Sprachentesten und -bewerten – auch international – weiterhin ein Desiderat (vgl. u. a. Harding & Kremmel 2016; Vogt & Tsagari 2014).

Aufgrund der mehrsprachigen Ausrichtung von IMoF finden sich in diesem Buch Beispiele aus allen Sprachen, die im Rahmen des IMoF unterrichtet werden: Englisch, Französisch, Italienisch, Russisch, Spanisch, Latein und Griechisch. Damit soll aber auch verdeutlicht werden, dass die Grundlagen und Prinzipien des Sprachentestens in gleichem Maße auf Fremdsprachen zutreffen. Die einzelnen Kapitel des Studienbuchs eignen sich zudem als theoretische Grundlage für Kurse in Fort- und Weiterbildungsveranstaltungen, die sich Themen der Leistungsmessung und -beurteilung widmen.

Wenn im Studienbuch der Einfachheit halber meist von ‚Tests' / ‚vom Testen' gesprochen wird, so sei an dieser Stelle angemerkt, dass damit unterschiedliche Formen der Leistungsüberprüfung gemeint sein können, wie etwa im schulischen Kontext Klassen- / Schularbei-

2 Nähere Informationen s. https://tinyurl.com/y7fkkvpp (21. 09. 2017).

ten, Klausuren, mündliche Prüfungen etc. Genauso können diese Begriffe aber auch für standardisierte nationale und internationale Tests stehen bzw. Abitur- und Reifeprüfungen meinen. Angemerkt sei, dass der *Gemeinsame europäische Referenzrahmen für Sprachen* (GeR) in seinem Untertitel neben den Tätigkeiten ‚lernen' und ‚lehren' auch von ‚beurteilen' (im Original: *learning, teaching, assessment*) spricht und in Kapitel 9 „Beurteilen und Bewerten" (im Englischen steht dafür der Begriff *assessment*) wesentliche Inhalte anführt, die „verschiedenen Funktionen des Prüfens und Beurteilens sowie entsprechenden Beurteilungs- und Bewertungsverfahren" gewidmet sind (Europarat 2001, 12). Auch in Kapitel 9 des GeR ist die Terminologie nicht einheitlich: ‚prüfen', ‚beurteilen', ‚bewerten' werden nebeneinander verwendet, genauso wie die Begriffe ‚testen', ‚prüfen', ‚bewerten', ‚beurteilen', ‚evaluieren' alltagssprachlich oft synonymisch verwendet werden, wenn z. B. von der Messung sprachlicher Kompetenz die Rede ist. ‚Testen' (*to test*) ist zweifellos der engste Begriff (*examining someone's knowledge*), ‚beurteilen' und ‚bewerten' (*to assess*) fassen die Tätigkeit der Leistungsüberprüfung weiter (*the goal of assessment is to make improvements*) und der Begriff ‚evaluieren' (*to evaluate*) hat die umfassendste Bedeutung (*making (institutional) judgements based on criteria and evidence*) (vgl. https://tinyurl.com/y82vcae2 [21. 09. 2017]).

Am Beginn eines jeden Kapitels im Buch finden sich Kann-Beschreibungen nach dem Muster des *Europäischen Portfolios für Sprachlehrende in Ausbildung* (EPOSA) (Newby et al. 2007), die einen Ausblick darüber geben, was den / die LeserIn im Kapitel erwartet, und die Ziele darlegen, wozu der / die LeserIn nach genauer Lektüre und Bearbeitung der am Ende eines jeden Kapitels angegebenen Arbeitsaufträge und Diskussionsfragen imstande sein soll. Die Tipps zu weiterführender Lektüre am Ende eines Kapitels dienen der Vertiefung der ausgeführten Inhalte und können genützt werden, um sich weiteres Wissen anzueignen. Die gesamte Literatur findet sich am Ende des Buches. Zudem sei an dieser Stelle auf den *Language Testing Bytes Podcast* verwiesen, in dem Glenn Fulcher begleitend zur Zeitschrift *Language Testing* aktuelle Fragen der Sprachtestforschung mit ExpertInnen diskutiert. Der Podcast erscheint halbjährlich und ist unter https://tinyurl.com/ycdpgjvr (21. 09. 2017) oder über *iTunes* verfügbar.

Marginalien am Textrand dienen der Strukturierung des Gelesenen; anhand dieser benutzerInnenorientierten Punkte kann sich der / die LeserIn – rekapitulierend in Form eines *self-assessment* – orientieren, ob er / sie die wichtigsten Inhalte eines Kapitels nachvollziehen und diese auch kurz erläutern kann.

Das Buch umfasst 11 Kapitel. Es wurde mit dem Ziel erstellt, auch im deutschsprachigen Raum ein Standardwerk zu „Testen und Bewerten fremdsprachlicher Kompetenzen" zur Verfügung zu haben, das gleichermaßen von Lehrenden und Lernenden an Universitäten sowie an Schulen genutzt werden kann, um die immer deutlicher eingeforderte „Bewertungskompetenz" einzelner *stakeholder* im Bereich fremdsprachlicher Leistungsmessung und -beurteilung zu stärken bzw. zu fördern.

In Kapitel 1 werden ein kurzer, historischer Überblick über die Entwicklung des Testens und Bewertens gegeben und drei Perioden des Sprachentestens vorgestellt, die als Beispiele für die Entwicklung von subjektiven, normorientierten Tests hin zu einer objektiven, validen und an Kriterien orientierten Bewertung dienen. In Kapitel 2 wird der GeR als kommunikativer, kompetenz- und handlungsorientierter Referenzrahmen des Europarates präsentiert, sein

Entstehungskontext beleuchtet, die Niveaustufen A1 bis C2 beschrieben und deren Bedeutung für das Testen und Bewerten von fremdsprachlichen Leistungen kritisch betrachtet. Die Hinwendung des GeR zu Sprachverwendenden als kommunikative, sozial Agierende und einer damit verbundenen positiven Sichtweise des Fehlers als inhärentes Kennzeichen von Lernersprache macht es erforderlich, die Rolle des Fehlers im Fremdsprachenunterricht neu zu überdenken, was in Kapitel 3 erfolgt.

In Kapitel 4 werden die Testgütekriterien in zwei Teilen vorgestellt: Im ersten Teil werden Arten der Objektivität, Reliabilität und Validität erklärt und beschrieben, wobei vor allem auf das zentrale Kriterium der Konstruktvalidität und den sich wandelnden Interpretationen der Validität bzw. des Prozesses der Validierung fokussiert wird. Im zweiten Teil wird auf die Prinzipien Authentizität, *Washback* und Praktikabilität eingegangen, der Bezug zwischen Testaufgaben und *real-world tasks* diskutiert, die Auswirkung von Tests auf Lehrende, Lernende, Unterricht und Bildungssystem illustriert sowie eine Kosten-Nutzen-Rechnung hinsichtlich Testressourcen aufgestellt. Der Testentwicklungszyklus wird in Kapitel 5 anhand von standardisierten Tests beschrieben; Begriffe wie Testzweck, Testarten, Testspezifikationen, *text mapping*, Prototypisierung, Pilotierung, Feldtestung, *Benchmarking* und *Standard-Setting* werden definiert und näher erklärt, um u. a. auf die hohe ethische Verantwortung im Bereich des Testens und Bewertens einzugehen.

Kapitel 6 widmet sich der Überprüfung rezeptiver Lese- und Hörverstehensleistungen. Anhand je eines konkreten Lese- (Nold & Willenberg) bzw. Hörverstehensmodells (Field) werden die einzelnen kognitiven Komponenten der nicht direkt beobachtbaren Leseverstehens- bzw. Hörverstehensprozesse aufgezeigt und vier prominente Lese- und Hörverstehensziele mit Bezug auf die GeR-Skalen erläutert. Es wird auf wesentliche Gemeinsamkeiten und Unterschiede bei der Überprüfung von Lese- und Hörverstehen hingewiesen, Testformate werden präsentiert, die sich zur Überprüfung eines Produktes, resultierend aus einer Lese- bzw. Hörverständnisaufgabe, eignen. Am Schluss steht ein Vorschlag, wie rezeptive Fertigkeiten als Basis für integrierte Testaufgaben genutzt werden können und welche Schwierigkeiten sich dadurch bei der Beurteilung ergeben.

In Kapitel 7 wird für die Beschreibung, wie produktive Fertigkeiten getestet werden können, ein ähnlicher Aufbau wie in Kapitel 6 gewählt. Das Konstrukt wird anhand je eines Modells (Shaw & Weir für Schreiben; Levelt für Sprechen) dargelegt, die GeR-Skalen für (monologische) Produktion und (dialogische) Interaktion werden in der Testanwendung konkretisiert. Außerdem werden Richtlinien vorgestellt für die Erstellung von lebensnahen, kontextualisierten und situationsgebundenen Testaufgaben mit unterschiedlichen Inputs (Texten, Bildern, Grafiken etc.) zur Überprüfung der Fertigkeit Schreiben (z. B. hinsichtlich des Einsatzes von Operatoren) bzw. für ein angemessenes InterlokutorInnen- respektive AssessorInnenverhalten bei der Überprüfung der Fertigkeit Sprechen. Dabei wird auf die Nutzung von holistischen und analytischen Bewertungsrastern im Sinne einer erhöhten Interrater-Reliabilität Bezug genommen; Vor- und Nachteile solcher Raster werden aufgezeigt.

Kapitel 8 widmet sich der Überprüfung sprachlicher Mittel in den linguistischen Kompetenzfeldern Lexik, Grammatik und Soziopragmatik. Grammatikalische Kompetenz wird als Teilkompetenz einer funktional-kommunikativen Kompetenz verstanden, für die angemessene

Testformate präsentiert werden. Lexik als wesentlicher Teil einer kommunikativen Verstehens- und Produktionsaktivität wird analog zum GeR einerseits mit Spektrum (Wortschatzbreite) und andererseits mit Beherrschung (Wortschatztiefe) assoziiert. In den Aufgabenformaten wird u. a. auf die Gebundenheit an einen Kontext (C-Test, *gap filling*) bzw. Losgelöstheit von einem Kontext (z. B. Übersetzungen) und die damit verbundenen Problematiken eingegangen. Soziopragmatische Kompetenz wird als wesentliche Komponente eines handlungsorientierten, kommunikativen und interkulturellen Fremdsprachenunterrichts erachtet, der im Unterricht mehr Bedeutung zukommen muss. Anhand des GeR wird an das wandelbare Konstrukt der Soziopragmatik angeknüpft, weil sich sprachliche und kulturelle Gegebenheiten in unserer globalen und digitalen Welt ständig verändern. Es werden Aufgabenformate zur Überprüfung soziopragmatischer Kompetenz unterbreitet, die in der Forschung Anwendung finden und für einen authentischen Einsatz in der Schule genützt werden können.

Kapitel 9 befasst sich für die klassischen Sprachen Latein und Griechisch mit der Überprüfung von Kompetenzen sowie deren sprachreflexiven Besonderheiten hinsichtlich der zentralen Fertigkeiten „Übersetzen" und „Interpretieren". Beides sind mehrstufige, komplexe Prozesse, die sowohl der Analyse als auch der Reflexion bedürfen. Bisherige Beurteilungs- und Korrekturpraktiken sorgten meist für negativen *Washback*, da „Sinn" als wichtigste Beurteilungsdimension schwer zu fassen und die bisherige Negativkorrektur der Validität nicht zuträglich war, sodass man dazu überging, objektivierbare Teilkompetenzen zu messen.

Kapitel 10 zeigt den komplexen Begriff der Beurteilungs- bzw. Bewertungskompetenz (*assessment literacy*) auf, der anhand der Bereiche *assessment of*, *assessment for* und *assessment as learning* näher beschrieben wird. In diesem Kapitel werden verschiedene Funktionen der Leistungsbeurteilung erläutert sowie alternative Formen der Beurteilung (wie *dynamic assessment*) oder Methoden zur Datenevaluierung wie *think alouds* vorgestellt, die eine Brücke zwischen Lehren, Lernen und Testen ermöglichen.

Das abschließende Kapitel 11 ist der, vor allem punktuellen, Leistungsbewertung im Schulalltag gewidmet und beleuchtet (in)formelle Tests und *teacher made tests*. Dabei wird der Frage nachgegangen, welche Aspekte Prüfungsaufgaben im schulischen Kontext aufweisen sollen, um Anforderungen wie Transparenz und gute Nachvollziehbarkeit zu erfüllen.

Abschließend sei folgenden Personen und Mitwirkenden aufrichtig und herzlich gedankt, ohne deren Unterstützung dieses Buch nicht möglich gewesen wäre: den AutorInnen der einzelnen Kapitel, Katrin Schmiderer für die professionelle und unermüdliche Arbeit am Manuskript, Herrn Seger, Frau Lembke und Frau Gastring vom Narr Verlag für ihre Geduld und die gute Zusammenarbeit, Margareth Graf und Renate Stadler für das aufmerksame Korrekturlesen und, *last but not least*, allen Studierenden, die die Ausbildung am IMoF durchlaufen haben und durch ihre kritischen Fragen, Anmerkungen und wertvollen Diskussionsbeiträge auch ImpulsgeberInnen für das vorliegende Buch waren.

Barbara Hinger und Wolfgang Stadler

1. Ein historischer Einblick in das Testen und Bewerten von Fremdsprachen

Barbara Hinger

Kann-Beschreibungen

Ich kann
- die historische Entwicklung des Sprachentestens in groben Zügen skizzieren.
- die drei Sprachtestparadigmen nach Spolsky (1976) erklären.
- aktuelle Desiderate der Sprachtestforschung beschreiben.

Die Forschungsliteratur zu Testen und Bewerten von Fremdsprachen kann bislang nur wenige Arbeiten nennen, die sich systematisch mit der geschichtlichen Entwicklung dieses Bereichs auseinandersetzen. Dabei verweisen die meisten AutorInnen zunächst auf die allgemeine Geschichte des Testens und Bewertens, die bereits in der Zeit der kaiserlichen Dynastien Chinas vor über 2000 Jahren, und damit sehr früh, einsetzte. Die damals etablierten Testverfahren dienten dem Zweck, die Bestqualifizierten – unabhängig von ihrer Zugehörigkeit zu einer bestimmten sozialen Klasse oder Familie – für den Staatsdienst auszuwählen (vgl. Spolsky 2008, 445; s. auch Fulcher 2010, 1 ff.; Kunnan 2008, 135; O'Sullivan 2012). Dieses Chinesische Prinzip (Macaulay 1853; Spolsky 1995) machte in anderen asiatischen Ländern, wie Korea oder Japan, ebenfalls Furore.

Normorientierte Bewertung bei der Auswahl der Besten nach dem Chinesischen Prinzip

Nach Europa gebracht wurde das Prinzip der Auswahl der Besten von den Jesuiten, die es geschickt mit dem hier im Mittelalter vorherrschenden Treviso-Prinzip (Spolsky 2008, 444) verbanden. Diesem ging es nicht um das Feststellen der Bestqualifizierten, sondern um den Nachweis der Leistung von SchülerInnen am Ende eines Lernjahres: Je nach Erfolg der SchülerInnen bezahlte die Stadt das Gehalt der verantwortlichen Lehrperson. Damit standen der curriculare Inhalt und dessen Umsetzung im Mittelpunkt: Erfüllten die SchülerInnen die Vorgaben zu den Lehrinhalten, hatten sie bestanden. Aus heutiger Sicht kann vermutet werden, hier einen Vorläufer kriterienorientierter, inhaltsvalider Verfahren vorzufinden, bei dem die Testkriterien auf dem Curriculum basieren und die gelehrten Inhalte mit jenen der Prüfungen übereinstimmen sollten. Demgegenüber wäre die chinesische Art des Überprüfens wohl als normorientiert zu charakterisieren: Die Leistung des Einzelnen wurde vermutlich zur Leistung der Gesamtheit der TestteilnehmerInnen in Beziehung gesetzt. War ein Jahrgang leistungsschwächer, konnte eine Person mittlerer Leistung eher zu den Besten zählen als in einem Jahrgang mit einer leistungsstarken Gruppe. Im weiteren Lauf der Geschichte bleiben beide Zugänge zum Testen und Bewerten erhalten. Sie finden sich auch in aktuellen Debatten und begleiten die Auseinander-

Treviso-Prinzip als Vorläufer kriterienorientierter Bewertung

setzungen insbesondere in Zeiten von Änderungen und Umbrüchen in einem Prüfsystem. Grundsätzlich ging es jedoch im Chinesischen Prinzip wie im Treviso-System darum, Günstlingswirtschaft durch Fähigkeits- und Leistungsnachweise zu ersetzen und damit einer subjektiv gehaltenen oder auf sozialen Faktoren beruhenden Auswahl eine Objektivierung der Leistungsbewertung gegenüber zu stellen. Diese zielte letztendlich auf Chancengleichheit ab (vgl. O'Sullivan 2012, 9). Historisch gesehen gelang es damit in China, den Einfluss der Aristokratie zurückzudrängen und eine kaisertreue Beamtenschaft zu etablieren (vgl. Kunnan 2008, 136). Auch das Auftreten einer *education industry*, die die verschiedenen Tests erstellte, war – inklusive negativer Rückkoppelungen (*Washback*) (s. Abschnitt 4.2.2) – schon zu beobachten (vgl. O'Sullivan 2012, 9 f.).

Aufgaben zur Überprüfung bestimmter sprachlicher Fertigkeiten waren in den chinesischen Tests bereits inkludiert. So musste nachgewiesen werden, dass man in der Lage war, einen politischen Essay zu schreiben oder Gedichte anhand formaler Vorgaben wie Reimbildung zu verfassen (vgl. Kunnan 2008, 136).

In Europa trugen vor allem die Universitäten zur Verbreitung von Tests und Prüfungen bei. Die Umgestaltung respektive Neuetablierung staatlicher Bildungssysteme, wie in Frankreich, Preußen und Österreich insbesondere im 18. Jahrhundert, und die damit einhergehende Ausweitung und Öffnung der Schulsysteme zogen ähnliche Effekte nach sich. Interessanterweise hinkte das britische System hier zeitlich gesehen hinterher, wie O'Sullivan ausführt:

Verbreitung von Tests und Prüfungen durch Universitäten und neu etablierte staatliche Bildungssysteme

> Testing became a bigger issue in Britain in the 19th century when the establishment realized they needed to select people according to capability and end the practice of patronage (the French and Germans had already come to that conclusion almost half a century earlier). The introduction of competitive examinations to the civil service in the UK was preceded by the Oxford University Commission, which led to the introduction of examinations within the education system in 1850, […]. (O'Sullivan 2012, 10)

In Großbritannien wurden Anfang des 20. Jahrhunderts Tests für Englisch als Fremdsprache für Personen eingeführt, die aus den Kolonien stammten und eine Ausbildung im britischen Bildungssystem anstrebten (vgl. O'Sullivan 2012, 11). In den USA reichen erste Vorläufer von *large-scale language tests* respektive Sprachtests für eine hohe Anzahl an TestteilnehmerInnen in die zweite Hälfte des 19. Jahrhunderts zurück (vgl. Kunnan 2008, 136 f.). Diese Sprachtests waren Kinder ihrer Zeit und nutzten Prüfformate, die die damals vorherrschende Fremdsprachenvermittlung, also die Grammatik-Übersetzungs-Methode, widerspiegelten. An dieser Art der Überprüfung von Sprache kam bereits früh Kritik auf, sodass neue Aufgabenformate wie ‚Richtig / Falsch'-, ‚Einfach- oder Mehrfachwahl'- und ‚Bemerke den Fehler'-Aufgaben entwickelt wurden (vgl. Kunnan 2008, 137), von denen man sich eine objektivere Beurteilung der Fremdsprachenkenntnisse erhoffte. Über-

Vorläufer von *large-scale language tests* ab der 2. Hälfte des 19. Jahrhunderts in den USA

setzungsaufgaben wurden dennoch beibehalten. Einen deutlichen Wendepunkt in der Geschichte des Sprachentestens setzte der Zweite Weltkrieg. Insbesondere in den USA wurde nun in einem großangelegten Programm, dem *Army Specialized Training Program*, wissenschaftlich an der Entwicklung von Sprachtests gefeilt (vgl. Kunnan 2008, 138). Diese Arbeit ging einher mit der Etablierung der Audiolingualen Methode als neuem Sprachlehr- und -lernansatz. Dieser war ebenfalls wissenschaftlich begründet und basierte auf einer engen Kooperation zwischen hochangesehenen Linguisten des Strukturalismus, wie Bloomberg und Fries, und exzellenten Psychologen der behavioristischen Schule, wie B. F. Skinner.

Zweiter Weltkrieg als Wendepunkt in der Entwicklung der Sprachtestung

In der Entwicklung des Testens und Bewertens von Sprache muss an dieser Stelle auf die erste Systematisierung der Geschichte von Sprachtests verwiesen werden, die von Spolsky (1976) vorgelegt wurde und uns gleichzeitig in die Gegenwart des Sprachentestens führt. Spolsky unterscheidet drei Perioden des Sprachentestens:

- das vorwissenschaftliche
- das psychometrisch-strukturalistische
- das psycholinguistisch-soziolinguistische Sprachtestparadigma

Drei Perioden des Sprachentestens

Diese Unterteilung kann einerseits als geschichtliche Entwicklung und damit als Abfolge auf globaler Ebene gesehen werden. Je nach lokal-nationalen Bedingungen können sich die drei Perioden andererseits aber auch überlappen und / oder gleichzeitig und nebeneinander existieren (vgl. Spolsky 1976, 11). Auch wenn Spolsky zum einen zwar darauf verweist, dass es sich bei seiner Einteilung um eine grobe Generalisierung handelt (vgl. ebd.), und er zum anderen mittlerweile von seiner zunächst getroffenen Einteilung mit sehr differenzierten Begründungen abrückt (vgl. Spolsky 2017), erscheint es im Folgenden doch nützlich, die Charakterisierung der drei Perioden etwas näher zu betrachten.

Das vorwissenschaftliche Sprachentesten zeichnet sich durch einen subjektiven Zugang zur Bewertung von sprachlichen Leistungen aus. Die Bewertung kommt ohne statistisch begründbare Auswertungsverfahren aus. Benotet wird die Sprachleistung beispielsweise anhand schriftlicher Performanzen der Lernenden oder nach einer kurzen mündlichen Äußerung. Sprachprüfungen liegen eindeutig in der Hand der Lehrpersonen und erfordern keine weitere Expertise: Wenn jemand eine Sprache lehren und unterrichten kann, dann wird davon ausgegangen, dass er / sie die Sprachleistungen der Lernenden auch bewerten kann (vgl. Spolsky 1976, 11 f.).

Subjektive Bewertung von mündlichen und schriftlichen Performanzen im vorwissenschaftlichen Sprachentest-Paradigma

Demgegenüber setzt die psychometrisch-strukturalistische Periode des Sprachentestens auf Expertentum. Nun gilt es, Sprachleistungen objektiv,

Möglichst objektive Bewertung vor allem rezeptiver Fertigkeiten durch geschlossene Aufgabenformate im psychometrisch-strukturalistischen Sprachtestparadigma

zuverlässig und wissenschaftlich begründbar zu überprüfen und zu bewerten. ExpertInnen in der Testtheorie sind verantwortlich für das Entwerfen adäquater Prüfformate und für deren statistische Auswertung, LinguistInnen geben die zu überprüfenden Sprachbereiche vor. Ausgangspunkt ist die Kritik an den zuvor subjektiv ausgerichteten Sprachprüfungen. So wird erstmals anhand von Untersuchungen gezeigt, dass die vorherrschende Bewertung schriftlicher Aufsätze subjektiv ausgeprägt und nicht reliabel ist (vgl. Hartog & Rhodes, 1936; Pilliner, 1952, zitiert in Spolsky 1976). Diesem Problem wird vor allem durch das Entwickeln geschlossener Aufgabenformate wie *multiple choice-* oder Einfachwahlaufgaben und halboffener Formate wie Kurzantworten versucht entgegenzuwirken, da deren Ergebnisse statistisch berechenbar sind und objektiv ausgewertet werden können. Damit wird der Fokus jedoch deutlich auf die Überprüfung der rezeptiven Fertigkeiten – Lesen, Hören – und der sprachlichen Mittel – Wortschatz, Grammatik – gelegt. Da die zu überprüfenden Sprachbereiche von der strukturalistischen Linguistik eingebracht werden, verwundert es nicht, dass diese auf der Basis kontrastiver Sprachvergleiche zwischen Ausgangs- und Zielsprache festgelegt werden und vor allem jene Strukturen überprüfen, die keine Gemeinsamkeiten in den betreffenden Sprachen aufweisen. Auf der Strecke bleiben eine umfassende Sicht von Sprache und ein adäquates Einbeziehen der produktiven Fertigkeiten Schreiben und Sprechen. Nichtsdestotrotz findet in der psychometrisch-strukturalistischen Sprachtestperiode die erste gezielte Zusammenarbeit zwischen den nach wie vor wesentlichen Bezugswissenschaften des Sprachenlernens, -lehrens und -testens, nämlich der Sprachwissenschaft und Psychologie, statt.

Geschlossene (vorgegebene Antwortmöglichkeiten), halb-offene (keine vorgegeben Antwortmöglichkeiten, erwartbare Antworten) und offene (keine vorgegebenen Antwortmöglichkeiten, freie Antworten) Aufgabenformate

Während Morrow (1979, 144) die erste Periode, also das vorwissenschaftliche Sprachentesten, metaphorisch als „Garten Eden" bezeichnet, in dem jeder / jede frei ist, zu tun und zu lassen, was ihm / ihr beliebt, nennt er die eben skizzierte psychometrisch-strukturalistische Sprachtestperiode das „Tal der Tränen": In diesem scheint alles reglementiert zu sein und die Messbarkeit überdeckt als wesentlichstes Ziel das tatsächliche Beherrschen und Sich-Ausdrücken-Können in einer Zielsprache.

Morrow zufolge wird mit der dritten Sprachtestperiode, dem psycholinguistisch-soziolinguistischen Sprachentesten, das „verheißene, gelobte Land" betreten. Nun rückt das Gütekriterium der Validität, also der Übereinstimmung zwischen einer umfassenden Konzeption von Sprache, wie sie im Unterricht vermittelt wird, und der Auffassung von Sprache, wie sie Sprachtests als theoretisches Konstrukt (s. Abschnitt 5.2) zugrunde liegt, in den Mittelpunkt. Sprachtestergebnisse sollen zwar weiterhin so objektiv und reliabel wie möglich sein, angestrebt wird nun aber, diese Kriterien auch auf die produktiven Sprachfertigkeiten zu übertragen. Dieses Unterfangen sollte beispielsweise durch die Bewertung von mündlichen oder schriftlichen Sprachleistungen anhand der Überprüfung festgelegter Kriterien gelingen. Diese kriterienorientierte Bewertung soll an die Stelle einer

Validität als wesentliches Testprinzip im psycholinguistisch-soziolinguistischen Sprachtestparadigma

subjektiven Notenvergabe treten und transparent gestaltet sein, indem die Bewertungskriterien auch den Lernenden zugänglich gemacht werden. Zudem sollte das Augenmerk auf Intra- und Interrater-Reliabilität gelegt werden (s. Abschnitt 4.1.2).

Sprachwissenschaftlich gesehen findet die Periode des psycholinguistisch-soziolinguistischen Sprachentestens in der sog. pragmalinguistischen Wende ihre Begründung. Diese setzt in den 60er Jahren des 20. Jahrhunderts ein und bedingt circa zehn Jahre später die kommunikative Wende im Fremdsprachenunterricht. Damit rücken die sprachliche Handlungs- und Kommunikationsfähigkeit in den Mittelpunkt des Unterrichts. Spolsky trägt diesem Paradigmenwechsel in Sprachwissenschaft und Sprachunterricht mit dem Adjektiv „soziolinguistisch" Rechnung. Die Bezeichnung „psycholinguistisch" lässt sich demgegenüber mit direkten und indirekten Auswirkungen des *cognitive turn* in der Sprachwissenschaft erklären. Dieser erlaubt insofern ein Abweichen vom Strukturalismus als Basis der Bewertung von sprachlichen Äußerungen, als er eine Grundlage für die empirische Auseinandersetzung mit realen Lerneräußerungen schafft und damit von einer kontrastiven Betrachtung sprachlicher Elemente in Ausgangs- und Zielsprache absieht. Die Betrachtung der tatsächlichen Sprachäußerungen von Lernenden wird nun postuliert und die Analyse der sich entwickelnden Lernersprache – *interlanguage* nach Selinker (1972) – ermöglicht.

Auf den Plan tritt somit die Psycholinguistik, die sich mit der mentalen Verarbeitung von Sprache beschäftigt. Auch wenn heute mittlerweile interessante theoretische Modellansätze vorliegen, sind wir nach wie vor weit von umfassenden, psycholinguistisch begründ- und beschreibbaren Entwicklungen des Sprachenlernens entfernt. Der Fremdsprachenunterricht kann also nur bedingt auf mögliche Handlungsanweisungen zurückgreifen, die lernersprachenbasiert sind (s. Kapitel 2 und 3). Die Forderungen, den Fremdsprachenunterricht und das Überprüfen von Leistungen in der Fremdsprache lernersprachensensibel auszurichten, werden jedoch immer stärker (vgl. u. a. Larsen-Freeman 2009; Van Moere 2012) und weisen in eine anstrebenswerte Richtung. Bei entsprechender Vorlage ausreichender empirischer Forschungsergebnisse aus der Spracherwerbs-, Sprachlehr- und Sprachtestforschung könnte Spolskys psycholinguistisches Paradigma des Sprachentestens auch erfüllt werden und möglicherweise den Zugang zum „gelobten Land", im Sinne Morrows, eröffnen, in dem Fremdsprachenunterricht und das Bewerten fremdsprachlicher Leistungen von Lernenden an einer realistischen Lernersprachentwicklung ausgerichtet sind.

Forderung nach lernersprachensensiblem Testen

In Anbetracht dessen erscheint es daher unter Einbeziehung des aktuellen Forschungsstandes adäquater, Spolskys dritte Periode des Sprachentestens nicht als „psycholinguistisch-soziolinguistisch", sondern als „kommunikativ-handlungsorientiert" zu bezeichnen. Damit kann auch auf die richtungsweisenden Sprachmodelle von Canale (1983), Canale & Swain (1980), Bachman (1990) sowie Bachman & Palmer (1996) verwiesen werden, die kommunikative Sprachkompetenzen umfassend definieren und Kriterien für ihre Überprüfbarkeit vorlegen. Das Einbeziehen aller sprachlichen Fertigkeiten und das Bemühen um eine adäquate Überprüfung von Wortschatz und Grammatik (s. Abschnitt 8.1) stehen aktuell im