In:『EUにおける言語問題と言語文化』平成１７年度～平成１９年度科学研究費補助金　課題番号17510203
基盤研究（C）研究成果報告書　古澤ゆう子代表者　２００８年

High-stakes-Sprachprüfungen in Japan, Washback-Effekt und der *Gemeinsame Europäische Referenzrahmen für Sprachen*

Ralph Degen

Einleitung

Dieser Aufsatz soll eine vorbereitende Untersuchung drüber sein, wie unvorteilhafte Auswirkungen von Sprachprüfungen in Japan, vor allem der Englischtests der Universitäts-Aufnahmeprüfungen, vor dem Begriff des Washback-Effekts[1] bewertet werten müssen, ob es möglich und sinnvoll erscheint die Fremdsprachenausbildung in Japan durch bewusst erzeugten positiven Washback zu verbessern und ob der *Gemeinsame Europäische Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*[2] (im weitern mit GER abgekürzt) dabei als erprobtes Werkzeug von Hilfe sein könnte.

Zu diesem Zweck soll zunächst der Begriff des negativen und positiven Washback in Bezug auf die Situation in Japan erörtert werden, wobei das Augenmerk vor allem auf das Gütekriterium der Validität gelenkt wird. Es wird daraufhin diskutiert, inwiefern der GER dazu beitragen könnte mit seinen handlungsorientierten Kann-Beschreibungen, die Validität der *high-stakes*-Tests[3] – vor allem des Englischteils der zentralen Aufnahmeprüfung, des sogenannten *sentâ-shiken* – deutlich zu verbessern, indem er einen Rahmen für ein kriterienbezogenes Konstrukt der Validität bietet, das darüber hinaus auch der Curriculumsbildung, Lehrmaterialerstellung und Lehrerausbildung zuträglich wäre.

Es sei darauf verwiesen, dass diese Untersuchung noch am Anfang steht und vor allem einen Umriss zukünftiger Forschung geben soll, die auch empirische Erhebungen beinhalten müsste. Das Ziel des Unterfangens ist es, die Testinhalte und -formate der *high-stakes*-Tests, die bisher offenbar zu stark nach dem Gesichtspunkt der Quantifizierbarkeit von Lernleistung zum Zwecke der Selektion gewählt werden, näher mit kommunikativer Kompetenz zu verbinden und dadurch erst die Bedingung für kommunikativen, handlungsorientierten Unterricht zu schaffen.

---

[1] Auf japanisch: *hakyû kôka* 波及効果. Während in der britischen angewandten Linguistik der Terminus „washback“ vorherrscht, wird oft auch der Begriff „backwash“ verwendet. (vgl. Alderson/Wall 1993, S. 115) Hier soll keine Unterscheidung zwischen den beiden Termini getroffen werden.

[2] Dokument im Auftrag des Europarates, dessen Inhalt weiter unten genauer erklärt wird. Titel des Englischen Originals: *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR). Titel der japanischen Übersetzung:『外国語の学習のためのヨーロッパ共通参照枠』. Der GER kann auf Deutsch, Englisch und Japanisch als Vollversion online gelesen werden: siehe Literaturangaben.

[3] Prüfungen, die von Lernenden, Lehrenden, Eltern und anderen Beteiligten als entscheidend für den weiteren Lebensweg (schulische oder berufliche Laufbahn) angesehen werden. (vgl. z.B. Madaus 1988, S. 87 ff.)

1. Washback

1.1. Begriffsdefinition

Der Begriff des Washback und seine Implikationen brauchen hier nicht ausführlich erörtert zu werden, dafür sei auf die einschlägige Literatur verwiesen (u. a. Alderson/Wall 1993 und 2001, Cheng 2005, Cheng/Curtis 2004, Wall 1996 und 2001). Es soll nur ein kurzer Überblick gegeben werden, da das Verständnis dieses Phänomens für den weiteren Gedankengang essenziell ist.

Cheng formuliert das Grundproblem so: „Traditionally, tests come at the end of the teaching and learning process. However, with the advent of high-stake public examinations testing nowadays, the direction seems to be reversed. Testing usually comes first before the teaching and learning process" (1997, S. 39, vgl. auch Pearson 1988, S. 98). Da Prüfungen zeitlich nach dem Unterrichten bzw. Lernen angesiedelt sind, ist vielen dieser Sachverhalt nicht bewusst (vgl. Popham 1987, S. 681). Da ein Test das prüfen soll, was ursprünglich das Lernziel war, ist es aber nachvollziehbar, dass die Testinhalte von der Planung her eigentlich am Anfang stehen. Optimalerweise, d. h. wenn ein Test valide ist, soll er genau das messen, was während des Lehr- und Lernprozesses das Lernziel war. Und da man sowohl beim autonomen Lernen als auch bei der Planung des Unterrichts oder eines Curriculums immer vom Lernziel ausgehen sollte, bedeutet dies, dass der Test inhaltlich das widerspiegelt, was am Anfang des Planungsprozesses steht.

Was der Begriff „Washback" außerdem aus bildungspolitischer Perspektive und vor allem in Hinblick auf *high-stakes*-Tests impliziert, ist die Dimension des Drucks, den er auf alle Beteiligten ausübt und durch den sowohl Lehre als auch Lernen stark beeinflusst werden.

Dieser Einfluss kann entweder positiv oder negativ ausfallen, wie in den folgenden Abschnitten näher erläutert werden soll. Nennenswerter Washback entsteht eigentlich nur durch *high-stakes*-Tests, deren Wesensmerkmal es ist, selektiv zu sein. Insofern sind solche Tests durchaus bildungspolitische Machtmittel, wie u. a. Madaus (1990) und Shohamy (2001) betonen.

1.2. Nachweisbarkeit

In der Literatur über Washback wird häufig das Fehlen empirischer Belege für die Existenz des Washback-Effekts[4] und die Art, wie er funktioniert bemängelt (vgl. u. a. Wall 1996, S. 338). Mittlerweile gibt es allerdings einige umfassendere, empirische Untersuchungen zum Washback-Effekt von *high-stakes*-Tests auch über einen längeren Zeitraum hinweg (vgl. u. a. Cheng 1997 und 2005, Shohamy et al. 1996). Das Gros allerdings sind eher punktuelle, kleine Erhebungen, die, wie in den betreffenden Aufsätzen meistens auch erwähnt wird, nur beschränkte Aussagekraft besitzen. Generell ist wohl zu sagen, dass es aufgrund der enormen Komplexität eigentlich unmöglich ist Washback detailliert und exakt nachzuweisen, geschweige denn zuverlässig zu prognostizieren. Dazu gibt es zu viele beteiligte Parteien und Faktoren, die sich gegenseitig beeinflussen. Und Tests probeweise auf einem eingeschränkten Gebiet,

[4] Bezeichnenderweise trägt einer der in der Literatur zum Thema Washback am meisten erwähnten Aufsätze den Titel „Does washback exist?" (Alderson/Wall 1993). Hier gibt es eine knappe Zusammenfassung des Artikels auf Japanisch: http://homepage3.nifty.com/xiaolin/Alderson,Wall.htm

zu implementieren, um den Washback isoliert zu beobachten, kann zu keinen aussagekräftigen Ergebnissen führen, da es sich bei solchen „Feldversuchen“ nicht um wirkliche *high-stakes*-Tests handelt. Es müsste also eine *baseline*-Studie durchgeführt werden und das Ergebnis optimalerweise noch durch einen weiteren Test empirisch validiert werden, wie es auch Wall vorschlägt: „Two of the problems involved in investigating this area are the need to compare groups before and after the introduction of a new test, and the need for an independent test which measures the ‘right’ things (the aims of the curriculum)“ (Wall 2000, S. 502). Hier stellt sich freilich auch die Frage, wie Ziele und Konsistenz des Curriculums, anhand derer die Validität des Tests gemessen werden soll, gerechtfertigt werden können. Darauf soll später noch eingegangen werden.

Zu den ständigen und durchaus auch berechtigten Forderungen nach empirischen Erhebungen zum Washback von *high-stakes*-Tests müssen zwei Punkte klargestellt werden. Erstens wird sich das Phänomen, wie bereits angedeutet, niemals vollständig quantifizieren lassen, wie es in den meisten empirischen Untersuchungen versucht wird, und zweitens kann es als evident bezeichnet werden und lässt sich einfach nicht bestreiten. Zumindest der negative Washback, der von Sprachwissenstests ausgeht, lässt sich auch ohne empirische Forschung festmachen und ist auch seit Jahr und Tag immer wieder beschrieben worden. Selbst Watanabe (1996, S. 319), der sich ansonsten mit definitiven Aussagen sehr zurückhält, weist darauf hin, dass sich der Washback-Effekt von *high-stakes*-Tests nicht mehr bestreiten lässt. Es bedarf sicher keiner empirischen Ergebung um auch zu dem Schluss zu kommen, dass dies definitiv der Fall ist, wenn man sich das Lernverhalten von Oberschülern, die Industrie, die mit den Aufnahmeprüfungen zusammenhängt, wie z.B. Verlage, die Vorbereitungsmaterialien herausgeben, *yobikô* (Vorbereitungsschulen) usw. ansieht.

Das Problem ist also sicherlich nicht, nachzuweisen, ob es einen Washback-Effekt überhaupt gibt, sondern zu untersuchen, wie er funktioniert und wie man ihn funktionalisieren kann, um positiven Washback zu erzeugen. Hier kann man auf Literatur über Bildungsinnovation *(educational innovation)* und Washback-Studien über *high-stakes*-Tests zurückgreifen, von denen es mittlerweile einige gibt.[5]

Versuchsläufe sind bei *high-stakes*-Tests wie oben erwähnt nicht möglich. Definitive Daten kann man nur bekommen, nachdem ein Test tatsächlich eingeführt wurde. Und da die Grundbedingungen von Land zu Land verschieden sind, lässt sich ein gewisses Maß an Ungewissheit wohl nie ganz ausräumen. Hinzu kommt, dass nicht immer entschieden werden kann, wodurch untersuchte Phänomene – ob positiv oder negativ – verursacht werden, ob sie wirklich dem Washback eines bestimmten Tests zugeordnet werden können oder anderen, z.B. kulturellen Gegebenheiten entspringen.

Wichtig ist allerdings auch, wie Alderson (1986, S. 95) betont, dass nachgewiesen werden muss, dass der Test, der durch eine Innovation ersetzt werden soll, nicht gut funktioniert oder sich durch einen

[5] Viele sind für Japan relativ irrelevant, wie etwa die Studien über Sri Lanka (vgl. Alderson/Wall 1993b), da dort die Voraussetzungen zu unterschiedlich sind. Andere sind sehr interessant und lassen durchaus Schlüsse auf Japan zu. So etwa die Untersuchungen über die Einführung des Certificate of Education Examination (HKCEE) in Hong Kong, wie bei Cheng (1997, 2004 und 2005) dokumentiert.

Es gibt auch einige Untersuchungen von Watanabe (1990, 1996, 2001 und 2004b) zu verschiedenen Aspekten des Washback der Aufnahmeprüfungen in Japan

negativen Washback-Effekt auszeichnet. Dieser ist in der Regel eigentlich einfach evident, kann aber durch empirische Erhebungen auch durch eine Analyse des Tests in Hinblick auf seine Konstruktvalidität bewerkstelligt werden. Auf methodische Aspekte der empirischen Untersuchung von Washback und ihre Kritik soll hier nicht weiter eingegangen werden.[6] Ebenso ist es wichtig, zu zeigen, dass ein Test, der einen anderen mit negativem Washback ersetzen soll, auch dazu geeignet ist, positive Auswirkungen zu zeitigen, indem er ein klares Konstrukt aufweisen kann, was getestet werden soll, und möglichst valide ist.

1.3. Negativer Washback

Grundsätzliche Problemlage

An dieser Stelle sollen die negativen Auswirkungen, die ein zentralisierter *high-stakes*-Sprachtest auf die Lerner und das Bildungssystem haben kann, soweit sie auf die Situation in Japan zutreffen, kurz dargestellt werden. Vornehmlich geht es dabei um die Aufnahmeprüfungen für die Universität, nämlich den sentâ-shiken und die Aufnahmeprüfungen der verschiedenen prestigeträchtigen Universitäten. Obwohl auch andere Tests, wie etwa Aufnahmeprüfungen zur Oberschule, schul- und universitätsinterne Prüfungen und Tests für andere Fremdsprachen im Rahmen der Erwachsenenbildung wie etwa das „Diplom Deutsch in Japan“ *(dokken)*[7] einen Washback-Effekt haben, ist dieser verglichen mit den Aufnahmeprüfungen eher gering und soll hier nicht extra behandelt werden.

Zunächst muss erwähnt werden, dass zentralisierte *high-stakes*-Tests deren Hauptfunktion die Selektion ist, ihrem Wesen nach grundsätzlich dazu neigen, eher negativen Washback zu erzeugen. Auf sie trifft zunächst die Kritik zu, die generell am traditionellen Ziffernnotensystem geübt wird: Die Noten (bzw. Punktzahlen) sind nicht vergleichbar und daher eigentlich nichtssagend, werden als Selektionsmittel missbraucht, beeinträchtigen die Kreativität der Lernenden und werden oft nicht den klassischen Gütekriterien von Validität, Reliabilität und Objektivität gerecht (vgl. Glaboniat 2006, S. 33). Ein Weiteres Problem ist die Stichprobenhaftigkeit dieser Art der Leistungsmessung, die vollständige Validität unmöglich macht, zudem sich viele schulische Leistungen nicht operationalisieren lassen. Deshalb kommt Beer zu dem Schluss: „Punktuelle Tests liefern im Zusammenhang mit Selektionsentscheidungen immer auch Falschzuordnungen“ (2006, S. 58).

Schon 1956 erwähnt Vernon „Excessive coaching for exams“ (nach Alderson/Wall 1993, S. 115), auf Deutsch also „exzessives Pauken“ als Ergebnis von Tests. Dadurch, dass in den meisten standardisierten *high-stakes*-Tests vor allem das einfacher messbare deklarative Sprachwissen statt kommunikativer Kompetenz, die zu messen einen deutlich höheren Aufwand bedeutet, geprüft wird, und Lehrer nur Pauken lassen, was im Test vorkommt und vernachlässigen, was nicht vorkommt, werden die Fertigkeiten und Inhalte, die in der Prüfung unterrepräsentiert sind, kaum gelernt. (vgl. Messick, S. 252) Oder wie es Bleyhl mit einiger Ironie formuliert: „Der traditionelle Fremdsprachenunterricht widerstand nicht immer der

[6] Hauptwerkzeuge sind hierbei Umfragen, Interviews und Unterrichtsbeobachtung. Zur Vertiefung siehe u. a. Cheng 2005, S. 62 ff. und Watanabe 2004a.

[7] 独検（ドイツ語技能検定試験）http://www.dokken.or.jp/

Versuchung, sich eine eigene Welt zu schaffen. *Man war dort versucht zu prüfen, was man messen kann, und lehrte, was man prüfen kann"* (2003, S. 38). Interessant ist, dass er das in der Vergangenheitsform schreibt. Messick und Bleyhl beschreiben damit im Grunde genau die Situation in Japan. Als konkretes Indiz können hier die Materialien aufgeführt werden, mit denen privat oder in den Vorbereitungsschulen für den *sentâ-shiken* gelernt wird. Es sind hauptsächlich Tests aus vergangen Jahren mit Kommentaren und Erklärungen *(mogi shiken)*, die durchgearbeitet werden. Diese beinhalten natürlich, wie auch der *sentâ-shiken* nur *marksheet*-Aufgaben und haben nichts mit offenen, sprachproduktiven, kommunikativen oder interaktiven Übungen zu tun. Wie stark dieser Einfluss sein kann, illustriert ein Beispiel von Pearson (1988, S. 102), der berichtet, dass in Sri Lanka der schreibproduktive Teil eines Tests im Unterricht einfach nicht behandelt wird, weil man den Test auch ohne ihn bestehen kann. Solange der Test, für den gelernt wird ein Sprachwissenstest ist, kommt es also zu einer Entfremdung des Sprachenlernens, indem es nur noch dem Ziel dient, einen selektiven Test zu bestehen, und das Moment der Sprachkompetenz völlig in den Hintergrund tritt. Dies lässt sich auch bei japanischen Studenten, die den Testlernmarathon gerade hinter sich haben, sehr deutlich beobachten. Viele lösen Aufgaben mithilfe von Grammatikkenntnissen, ohne auch nur auf die Idee zu kommen, den Inhalt der Sätze zu berücksichtigen, die sie da lesen oder schreiben. Grundsätzlich werden hierdurch gesprochene Sprache und Sprachproduktive Fähigkeiten vernachlässigt. Kikuchi (2006, S. 94) weißt in seiner Analyse des sentâ-shiken und von Aufnahmeprüfungen von jeweils zehn privaten und staatlichen, bzw. öffentlichen Universitäten auf das Validitätsproblem hin, dass bei den Aufnahmeprüfungen eher die Fähigkeit, Testaufgaben zu lösen *(test-taking ability),* als Sprachkompetenz *(proficiency)* gemessen wird.[8]

Eine Episode Sekiguchis illustriert diese Diskrepanz von Wissen und Können, die durch das exzessive Testlernen deklarativen Sprachwissens entsteht, sehr gut. Er wurde im Goethe-Institut in Deutschland nach einem schriftlichen Einstufungstest in den höchsten Kurs eingeordnet, weil der den Einstufungstest perfekt gelöst hatte. Zu diesem Zeitpunkt war er schon einige Jahre Assistenzprofessor und Doktorand an der Keiô-Universität. Die Testaufgaben waren in etwa dieselben, die er immer seinen Studenten gab. In der ersten Stunde blamierte er sich, weil er völlig eingeschüchtert von der Sprachkompetenz der anderen Kursteilnehmer kein Wort über die Lippen brachte. Daraufhin wurde er in das Büro des Lehrers gerufen, der ihn fragte, wie er sich seine Zukunft als Germanist vorstelle, wenn er kein Deutsch könne (vgl. Sekiguchi 2000*,* S. 54).

Krumm spricht in diesem Zusammenhang vom Problem der Orientierung des Unterrichts auf den Output, der in einem standardisierten Test gemessen wird. „Die Sprachlehrforschung ebenso wie die moderne Mehrsprachigkeitsforschung betonen zu Recht immer wieder, dass ein wichtiges Ergebnis fremdsprachlichen Lernens der nicht unbedingt als Output messbare Erfahrungsgewinn in prozessualer und

[8] In der Studie folgt er einer Untersuchung von 1995, in der die Autoren von „test-wiseness" (nach Kikuchi 2006, S. 94) sprechen. Er kommt zu dem Schluss, dass sich die Aufnahmeprüfungen zwischen 1994 und 2004 nicht grundlegend geändert haben.

2006 wurde beim sentâ-shiken zwar ein Hörverstehensteil für Englisch aufgenommen, dieser ist im Vergleich zum schriftlichen Teil, vor allem zum Leseteil, allerdings unverhältnismäßig einfach.

persönlichkeitsbildender Hinsicht ist, bei dem auch die außerunterrichtlichen Lernprozesse (Lektüre, Recherche im Internet, Austausch und Begegnung) eine wichtige Rolle spielen“ (2006, S. 31/32). Dies führt seiner Meinung nach zur Verhinderung einer spezifischen Lernkultur und individueller Lernprofile (vgl. ebd.)

Testlernen hat auch zur Folge, dass vor dem Test Gelerntes nach dem Test vernachlässigt wird, um sich auf den Inhalt des nächsten Tests vorzubreiten. Hinzu kommt, dass nicht verinnerlichtes Sprachwissen auch sehr schnell wieder vergessen wird. Es wäre also notwenig, Tests und Unterricht so anzulegen, dass bereits gelerntes im weiteren Unterricht vertieft und nicht vergessen wird, wie es auch im GER betont wird „Jedes Niveau sollte so verstanden werden, dass es alle anderen, niedrigeren Niveaus auf der Skala mit einschließt. Das heißt, dass jemand, der B1 (*Threshold*) erreicht hat, auch alles tun kann, was in A2 (*Waystage*) aufgeführt ist und dies sogar besser kann, als in A2 beschrieben.“ (GER, Kap. 3.7)[9] Das sollte sich auch in der Test- und Unterrichtsplanung widerspiegeln, um negativen Washback zu vermeiden.

Aus der Sicht der Testtheorie lassen sich die Ursachen für die oben dargestellte Diskrepanz zwischen deklarativem Sprachwissen und kommunikativer Kompetenz mit den Begriffen der „Unterpräsentation des Konstrukts“ *(construct underrepresentation)* und der „konstruktirrelevanten Schwierigkeit“ *(construct-irrelevant difficultiy* oder *variance)* beschreiben (vgl. Messick 1996 und 1994). Sie sollen helfen zu bestimmten, wie Tests aussehen, die vermutlich einen negativen Washback erzeugen und solche, die einen positiven Washback erzeugen.

Unterrepräsentation des Konstrukts

„Wie aber soll die Sprache, die […] die komplexeste Erfindung des Menschen ist, selbst bei einem Lerner in ihrer auch dort gegebenen Vieldimensionalität vermessen werden? Nun gibt es einen Bereich, wo es den Anschein hat, dass am ehesten Konsens oder Objektivität erreicht werden könnte, nämlich bei der formalen Dimension. Das Problem für den Sprachlehrer besteht jedoch darin, dass beim Anlegen der Elle zur Qualitätsmessung an dieser Stelle just die Gefahr der lernpsychologischen Kontraproduktivität am größten ist. Das Ziel einer Sprachkompetenz in einer Fremden Sprache – oder gar in mehreren – mittels eines streng formorientierten Unterrichts zu erreichen, ist damit – so lehrt die Erfahrung – am allerwenigsten zu erreichen“ (Bleyhl 2003, S. 36) Mit dieser Aussage fasst Bleyhl das Problem der Unterrepräsentation des Konstrukts und seine Konsequenzen komprimiert zusammen. Wie bereits oben erwähnt, orientieren sich Testentwickler (auch Lehrer usw.) bei der Gestaltung von Sprachprüfungen nicht in erster Linie an den Fähigkeiten und Fertigkeiten, die notwendig sind, um sprachliche Kompetenz zu entwickeln, sondern der Fokus verrutscht in Richtung einfacher Messbarkeit. Dadurch wird das Konstrukt, bei formalem, deklarativem Sprachwissen, überrepräsentiert. Dies geschieht auf Kosten anderer Fähigkeiten, die ebenso notwenig sind und daher auch gemessen werden müssten, um durch einen Test ein korrektes Bild der

[9] Zitate aus dem GER sind mit Kapitel- und nicht mit Seitenangaben gekennzeichnet, weil nicht auf die Druck- sondern auf die Onlineausgabe zurückgegriffen wurde: http://www.goethe.de/Z/50/commeuro/

Sprachkompetenz eines Testnehmers zu ermitteln. Oder wie Glaboniat es ausdrückt: „In den Bereich der Validität fällt aber auch, dass die Beurteilung möglichst alle Kompetenzbereiche abdeckt“ (2006, S. 45).

Das Problem bei den Aufnahmeprüfungen und auch bei anderen Tests in Japan scheint allerdings noch tiefer zu sitzen. Die Selektionsfunktion der Aufnahmeprüfungen ist so stark in den Vordergrund gerückt, dass sich die Frage stellt, ob dem Test überhaupt ein klares Konstrukt zugrunde liegt. Wie dem auch sei, es ist wohl offensichtlich, dass der sentâ-shiken und die Aufnahmeprüfungen, zumindest der zwanzig von Kikuchi (2006) untersuchten Prestigeuniversitäten ein Validitätsproblem haben. Dies wirft allerdings ein weiters Problem auf. „Bei der Validität handelt es sich hinsichtlich der Testpraxis um das wichtigste Gütekriterium überhaupt. Die Gütekriterien Objektivität und Reliabilität ermöglichen eine hohe Messgenauigkeit, liefern aber nur die günstigen Voraussetzungen für das Erreichen einer hohen Validität“ (Schermelleh-Engel, S. 4). Die vermeintliche Objektivität und Reliabilität der Aufnahmeprüfungen, vor allem des sentâ-shiken können also nicht wirklich als hinreichend bezeichnet werden, weil Objektivität und Reliabilität nur gegeben sind, wenn ein Test valide ist. Aus dem Verhältnis der drei Gütekriterien, die sich gegenseitig bedingen, geht also zum einen hervor, dass ein Test, der nicht objektiv und reliabel ist, auch nicht als valide bezeichnet werden kann. Andererseits ist Validität aber eine notwendige Bedingung für Objektivität und Reliabilität (vgl. Grotjahn 2000, 312 ff.). Konkret kann dies wie folgt veranschaulicht werden: „if what is under-represented in the assessment of communicative competence is an important part of the criterion performance, such as listening and speaking as opposed to reading and writing, then invalidly high scores may be attained by examinees well prepared on the presented skills but ill prepared on the underrepresented ones“ (Messick 1996, S. 252). Washback bedeutet, dass sich dieses Ungleichgewicht im Unterricht und im Verhalten der Lerner, Lehrer und anderer Beteiligter niederschlägt.

Ein weiterer Aspekt ist, dass beim Testen im traditionellen Sinne die Fertigkeiten so weit wie möglich getrennt geprüft werden. Dabei wäre es weniger schädigend, wenn tatsächlich kommunikative Kompetenz, also die Verwendung mehrer Fertigkeiten geprüft würde.

Zunächst müsste also ein klar formuliertes Konstrukt, ein Lernziel ins Spiel kommen. Hierbei könnten Leistungsstandards die auf dem GER basieren sicherlich eine große Hilfe sein, wie Glaboniat (2006, S. 45) vorschlägt.

Konstruktirrelevante Schwierigkeit (construct-irrelevant difficultiy)

Mit konstruktirrelevanter Schwierigkeit bezeichnet man durch das Testformat geschaffene Schwierigkeiten, die mit der eigentlichen kommunikativen Kompetenz auf einem bestimmten Niveau, auf dem getestet werden soll, nichts zu tun haben. Beispiele dafür wären im Grunde alle Tätigkeiten, Fertigkeiten und Kenntnisse, die in einer tatsächlichen kommunikativen Situation nicht vorkommen bzw. nicht gebraucht werden, in der Testsituation aber gefordert werden. Dazu kann etwa die exzessive Verwendung von Aufgabenformaten, die in realen Situationen der Sprachanwendung nicht vorkommen, wie z.B.

Einsetzübungen, *multiple-choice*-Aufgaben usw. oder dem eigentlich getesteten Niveau unverhältnismäßig schwierige Grammatik gezählt werden.

Auch Übersetzen auf niedrigen und mittleren Niveaustufen muss dazugerechnet werden. Nicht zu Unrecht wird Übersetzen im GER erst auf der höchsten Niveaustufe erwähnt. Im Unterricht nach der Grammatik-Übersetzungs-Methode (GÜM) kommt es wohl hauptsächlich als Instrument zur Kontrolle, ob die Lerner einen Text verstanden haben, vor. Wie man bei den Aufnahmeprüfungen in Japan sieht, ist das Übersetzen auch ein beliebtes Testformat mit starkem negativem Washback (vgl. Watanabe 1996). Die irrelevante Schwierigkeit liegt darin, dass schon auf niedrigem Niveau alles detailliert gelesen wird, während globales, selektives und extensives Lesen vernachlässigt oder überhaupt nicht berücksichtigt werden. Dabei hat detailliertes Lesen auf niedrigen Niveaustufen am wenigsten zu suchen. Umfragen bei Studierenden an japanischen Universitäten zeigen auch tatsächlich, dass nur die wenigsten über ein breites Spektrum an Lesestrategien verfügen und die meisten auch noch nie einen längern Text auf Englisch gelesen haben. Wie Komárek erwähnt, können sich zu häufige Übersetzungsübungen im Unterricht sogar kontraproduktiv auswirken, weil die Lerner dann „ständig zwischen Ziel- und Muttersprache hin und her springen“ (2006, S. 11), anstatt die Fremdsprache als eigenes Bezugssystem aufzubauen und zu verwenden, wie es nach Reiss (1983) erfolgreiche Sprachenlerner machen. Auch hier lassen sich die Auswirkungen bei japanischen Studenten leicht auffinden.

Auswirkung auf die Lerner

Abgesehen davon, dass Lerner, die hauptsächlich auf einen *high-stakes*-Tests hin „pauken“ und unterrichtet werden, zumindest im Verhältnis zum Lernaufwand, nur geringe Sprachkompetenz entwickeln, zeichnet sich der Washback-Effekt noch durch eine Reihe anderer Langzeitwirkungen aus, die nicht übersehen werden dürfen.

Eines der wohl auffälligsten Merkmale ist die Passivität. Nakajima (1997) spricht von „Horden schweigender Studenten“. Man kann sicherlich nicht sagen und es wäre auch schwer nachzuweisen, dass diese Passivität alleine vom Washback-Effekt der Aufnahmeprüfungen herrührt, da es natürlich auch kulturelle Implikationen gibt, die Passivität von Schülern und Studenten befördern (vgl. hierzu ebenfalls Nakajima 1997). Andererseits ist es auch nicht anders zu erwarten, als dass Unterricht, der hauptsächlich testrelevantes Sprachwissen vermittelt diesen Effekt hat. Generell kann eigentlich nur handlungsbezogener Unterricht, in dem hauptsächlich Aufgaben vorkommen, die prozedurales Wissen fördern, Lerner zu aktivem Unterrichtsverhalten animieren. Die ausschließliche Vermittlung deklarativen Wissens muss zwangsläufig zu Passivität führen. „Fertigkeiten und prozedurales Wissen […] basieren mehr auf der Fähigkeit, Handlungen und Prozesse auszuführen als auf deklarativem Wissen, obgleich solche Fähigkeiten durch den Erwerb von "vergessbarem" deklarativem Wissen gefördert werden können.“ (GER, Kap. 2.1.1)

Es ist auch eine sehr starke Abhängigkeit der Lerner zu erkennen, die zwar zum Teil durch

Lehrtraditionen[10] in Japan begünstigt wird, aber auch auf Washback zurückgeführt werden kann. Die Aussage einer Deutschlehrerin in England illustriert einen ähnlichen Sachverhalt: „Die Lehrer machten geradezu alles für die Schüler. Für alle im Examen geforderten Redesituationen bereiteten sie Frage-Antwort-Listen vor, die die Schüler auswendig lernen sollten, manchmal ohne wirklich zu verstehen, worum es wirklich ging." (Butzkamm/Plum 2006, S. 37). Selbständiges, forschendes Lernen wird so unterbunden, und durch ein Reiz-Reaktionsprinzip mit vorgegebenen Lösungen auf vorgegebene Aufgaben reduziert. Metakognitive Fähigkeiten können sich auf diese Weise kaum entwickeln, was aus Umfragen mit Studenten und zahllosen Erfahrungen im Unterricht klar hervorgeht. Wenden (1991, S. 57) verwendet hier den aus der Psychologie entlehnten Begriff „learned helplessness": Studenten betrachten sich selbst als unfähig, ohne detaillierte Anweisung durch den Lehrer Sprachen zu lernen. „Some university students often ask how they should learn EFL in a better way. These students, who have passed the entrance examination, seem to be lost when it comes to learning EFL efficiently in an examination-free situation" (Watanabe 1990, S. 176)[11]

Das Primat der Grammatik, also das Hauptgewicht auf Aspekten formaler Korrektheit darf in diesem Zusammenhang nicht unerwähnt bleiben. Induktives Erarbeiten Grammatischer Strukturen kommt offenbar sehr selten vor. Es wirkt eher störend, weil es nicht relevant für den begrenzten Stoff ist, der abgefragt wird und mehr Zeit in Anspruch nimmt, als das Auswendiglernen vorgegebener Patterns. Da prozedurales Wissen, das viel schwieriger durch objektive Testaufgaben zu quantifizieren ist, im Unterricht fast überhaupt nicht vorgesehen ist, bleibt nur deklaratives Wissen. Lernen unter dem Primat der Grammatik ist kontraproduktiv. Es führt zu „didaktischer Überbehütung", die natürliches Lernverhalten behindert. „Grammatik kann nicht gelehrt werden, sie muss vom Lerner entdeckt werden" (Bleyhl 2003, S. 40). Dabei ist zu beachten, dass sich Lehr- und Lernverhalten perpetuieren. Lehrer unterrichten, wie sie unterrichtet wurden, auch wenn kein *high-stakes*-Tests mehr ansteht, und Lerner neigen oft dazu, sich Tests zu suchen, für die sie lernen können.

Die übermäßige Gewichtung formalsprachlicher Korrektheit in Hinblick auf Punktabzüge bei Fehlern in der Prüfung hat den Effekt, dass eine Interimssprache weder im Bewusstsein der Lerner noch des Lehrpersonals zu existieren scheint. Ergebnis sind ständige Korrekturen und der daraus resultierende Verlust sprachlicher Spontaneität, der letztlich im bereits erwähnten Schweigen im Unterricht gipfelt. Diese Fehlerangst sitzt tief und wirkt lange nach, oft ein ganzes Leben. Die Verbindung aus Angst vor Fehlern, mangelnden Kommunikationsstrategien und dem Selbstbewusstsein unzureichender sprachlicher Kompetenz wiederum führt zu Furcht, die gelernte Fremdsprache „im richtigen Leben" anzuwenden. „Die meisten Schüler wissen wohl, dass sie im Grunde nichts können und haben eine Heidenangst, wenn sie frei sprechen sollen. Einige haben dann auch Bammel davor, nach Deutschland zu fahren" (Butzkamm/Plum

[10] Es ist, wie bereits angedeutet, kaum möglich Phänomene definitiv Washback oder anderen, vornehmlich kulturellen Einflüssen zuzuordnen, zumal diese sich auch gegenseitig beeinflussen. Zumindest ein erheblicher Einfluss kann allerdings auf Washback zurückgeführt werden, weil dieselben Phänomene auch in anderen Ländern zu Tage treten und kausale Verbindungen offensichtlich sind.

[11] Zum Thema Lernerautonomie bei japanischen Studenten siehe auch Degen 2004.

2006, S. 37). Das ist in Japan übrigens nicht nur bei Lernern, sondern auch bei Sprachlehrern zu beobachten. Tornberg spricht in diesem Zusammenhang von Abbau von Selbstvertrauen durch Bestrafung mit schlechten Noten, obwohl die Schüler gar nicht wissen, wofür sie eigentlich lernen. (1996, S. 24). Aus lernpsychologischer Sicht sei noch erwähnt, dass sich Fehler- und Prüfungsangst durchaus hindernd auf den Lernprozess auswirken können.

Als letztes muss noch der Washback-Effekt auf die Motivation von Lernern erwähnt werden. Die drastische Schilderung von Berwick und Ross trifft vermutlich nicht auf alle Studenten so zu, stimmt aber tendenziell. „At the same time [the last year of high school], the focus of the English examination is on grammar and translation […]. **Motivation** to learn English is thus channeled into the sort of proficiency with the least communicative value. Once the university examinations are over, there is very little to sustain this kind of motivation, so the student appears in freshmen classrooms as a kind of timid, exam-worn survivor with no apparent academic purpose at university“ (1989, S. 206). Bei vielen Lernern lässt sich tatsächlich eine deutliche Abneigung gegen Sprachenlernen und Frustration wegen mangelnder kommunikativer Kompetenz feststellen.

Es wird oft behauptet, dass sich *high-stakes*-Tests durch ihre subjektiv wahrgenommene Wichtigkeit motivierend auswirken. Tatsächlich wenden Testanwärter ja auch viel Zeit und Energie auf. Allerdings handelt es sich dabei zunächst nicht um intrinsische Motivation, sondern um externe, die sich verflüchtigt, sobald der Test bestanden ist. Auch Watanabe räumt ein, dass die Englische Sprache von vielen Schüler nur als Mittel betrachtet wird, an die Universität ihres Wunsches zu kommen: „It appears then that the washback effects of the entrance examination, which pervaded high school education, drove students to learn EFL only in order to pass the examination“ (1990, S. 188). Dies ist im Grunde auch nachvollziehbar, wenn man bedenkt, dass das Lernen nicht den eigenen Lernzielen gefolgt ist, sondern in Vorbereitung auf Tests geschah, die wenig Augenscheinvalidität besitzen.

1.4. Positiver Washback

Die Grundannahme, die hinter der Idee des Washback steht, formuliert Messick wie folgt: „a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning. Some proponents have even maintained that a test's validity should be appraised by the degree to which it manifests positive or negative washback" (1996, S. 241). Frederiksen und Collins benutzen den Begriff „systematic validity“: „A systemically valid test is one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time“ (Frederiksen/Collins 1990, S. 4/5). Das Ziel ist es also, das Lehrverhalten methodisch und inhaltlich zu verändern (vgl. Alderson/Banerjee 2001 S. 214), bzw. wie im Falle Japans, wo es bereits einen dominanten *high-stakes*-Test gibt, der, wie oben erwähnt, einen starken negativen Washback erzeugt, die Bedingungen neu zu setzen, unter denen Unterricht erst möglich wird, dessen Ziel tatsächlich Sprachkompetenz ist.

Folglich liegt der Knackpunkt zwischen negativem und positivem Washback, wie es auch Pearson hervorhebt, darin, ob ein Test handlungsorientiert ist oder Sprachwissen testet: „However one distinction that seems to be of special significance is that between tests of performance und tests of knowledge“ (Pearson 1988, S. 102 ff.)

Man kann generell sagen, dass Tests das Potential für einen positiven Washback haben, wenn man dieselben Aktivitäten, die man während des Tests ausführt, auch als Unterrichtsaktivitäten verwenden kann und sie modernen Ansprüchen an kommunikativen Unterricht genügen. Ebenso kann man sagen, dass viele Übungen und Unterrichtsaktivitäten auch als Testaufgaben verwendet werden können (vgl. Pearson 1988, S. 106 ff.).

Validität und Sprachstandards

Watanabe kommt in seinem Aufsatz über die Lehrerrolle im Zusammenhang mit dem Washback der Aufnahmeprüfungen in Japan zu einem sonderbaren Schluss: “It seems to be crucial then to identify empirically a range of effective teaching methods to improve authentic or real life language skills as well as to help students pass the examination” (2004, S. 140). Damit will er wohl zum Ausdruck bringen, dass es die Aufgabe von Lehrern ist, den Unterricht so zu gestalten, dass die Schüler auf die Aufnahmeprüfungen vorbereitet werden, gleichzeitig aber auch kommunikativen Unterricht zu machen, in dem die Schüler kommunikative Sprachkompetenz erwerben können. Momentan bleibt Lehrern, sofern sie denn fähig und daran interessiert sind, kommunikativen und handlungsorientierten Unterricht zu machen, auch nichts anders übrig. Vielversprechend kann dieser Kompromiss auf keinen Fall sein, weil die Notwendigkeit, den Test zu bestehen von den Lernen (und den meisten Lehrern sicherlich auch) als viel stärker empfunden wird, als der vage Wunsch die Fremdsprache zu beherrschen. Viel sinnvoller wäre es natürlich, wenn sich die Unterrichts- und Lernaktivitäten beim Testlernen mit dem Lernziel kommunikativer Sprachkompetenz decken würden. Solange diese Bedingung nicht erfüllt ist, erscheint es unmöglich, vor den Aufnahmeprüfungen wirklich effizienten Fremdsprachenunterricht zu machen. Diese Notwendigkeit führt direkt zur Forderung nach höherer Validität der Aufnahmeprüfungen, sowohl des sentâ-shiken als auch der universitätseigenen, in Bezug auf Handlungsorientierung und kommunikative Kompetenz**.**

Kriterienbezogene Validität der *high-stakes*-Tests nennt Popham auch als wichtige Bedingung für die Erzeugung positiven Washbacks.[12] „[…] the descriptive clarity of well-constructed criterion-referenced tests gives teachers comprehensible descriptions of what is being tested“ (Popham 1987, S. 680) Nur dadurch lässt sich auch die notwendige Transparenz erzeugen, wie auch Glaboniat betont: „Ein wesentlicher Ausgangspunkt für aussagekräftigere und validere Beurteilungsverfahren ist ein sachorientiertes (kriterienorientiertes) Vorgehen. Erst wenn die Prüfungsinhalte bzw. Grundlagen der Beurteilung transparent festgelegt und für alle einsehbar sind, kann eine Note ihre Berichts- und Selektionsfunktion erfüllen“ (Glaboniat 2006, S. 50). Dabei betont sie, dass kriterienbezogene Validität und

[12] Sein Terminus dafür ist *measurement-driven instruction* (MDI). Kriterienbezogene Tests *(criterion-referenced tests)* im Gegensatz zu normbezogenen Tests, bei denen die Leistungen von Lernern in Relation zueinander gemessen werden.

die daraus gewonnene Transparenz durchaus auch eine Bedingung für die Selektionsfunktion bei Tests ist. Ist diese nämlich nicht gewährleistet, ist auch nicht klar, auf welcher Basis ein Testnehmer besser bewertet wird als ein anderer. Außerdem sind Noten bzw. Punktzahlen ohne kriterienbezogene (sachbezogene)[13] Normen außerhalb der Bezugsgruppe aussagelos. „Transparenz und Vergleichbarkeit ist also nur möglich, wenn das, was überprüft werden soll, vorher festgelegt wird und eine kriteriumsorientierte Beurteilung erfolgen kann“ (Glaboniat 2006, S. 35).

Messick fordert daher, dass Tests authentische und direkte Beispiele kommunikativen Verhaltens aller vier Teilfertigkeiten in der Zielsprache enthalten sollen und Testaufgaben nahtlos in Übungsaufgaben übergehen sollen, so dass Testvorbereitung und Lernen der Sprache möglichst dieselben Aktivitäten sind (Messick 1996, S. 241/242, vgl. auch Andrews 2004, S. 49). Dass zentralisierte *high-stakes*-Tests vollkommen den Anforderungen handlungsorientierten Spracherwerbs genügen, also alle Aspekte kommunikativer Kompetenz erfassen, wird natürlich nie möglich sein, weil sich nicht alle Aspekte des Lernens eindeutig prüfen lassen und weil der Aufwand schlichtweg zu hoch wäre. Es ist also ein Kompromiss gefordert, der möglichst nah an tatsächliche Sprachkompetenz herankommt. Zu diesem Zweck muss negativer Washback minimiert werden, indem die Quellen von Invalidität so weit es psychometrisch und logistisch möglich ist, beseitigt werden, nämlich Unterrepräsentation des Konstrukts und konstruktirrelevante Schwierigkeiten, auf die oben bereits eingegangen wurde.[14]

Zur Erstellung eins Konstrukts handlungsorientierter, kommunikativer Sprachkompetenz sind Sprachstandards notwendig, die diesen Zielen gerecht werden: „Grundlage jeglicher Leistungs- bzw. Bildungsstandards müssen wissenschaftlich fundierte und fachdidaktisch akzeptierte Kompetenzmodelle sein“ (Glaboniat 2006, S. 40). Genau hier kann der GER ein zuverlässiges Werkzeug sein, da er, wie es seine Autoren explizit sagen, diesen Zweck erfüllen soll: „Der *Referenzrahmen* versucht, eine solche Basis für die **Beschreibung der Inhalte** und einen Fundus für die **Entwicklung genau definierter, spezifischer Kriterien** für direkte Tests zur Verfügung zu stellen.“ (GER, Kap. 9.3.8, Hervorhebungen im Original). Auf die Funktion des GER wird weiter unten genauer eingegangen.

Bewusst positiven Washback erzeugen

Die Idee, den Unterricht durch die Implementierung oder Veränderung von *high-stakes*-Test und dem daraus resultierenden positiven Washback-Effekt zu verbessern ist durchaus nicht neu. Popham verwendet in diesem Zusammenhang das Schlagwort der *measurement-driven instruction* (MDI, vgl. Popham 1987). In Anlehnung daran betonen Davidson und Fulcher den utilitaristischen Aspekt dieses Ansatzes: „but what

[13] Man sagt auch „kriteriumsorientiert“ bzw. „lernzielorientiert“ in Abhebung zu „normorientiert“ bzw. „bezugsgruppen-orientiert“ (vgl. Grotjahn 2000, S. 329 ff.)

[14] In seinem Aufsatz „Validity and washback in language testing“ gibt Messick auch einen Fragenkatalog mit dessen Hilfe Störfaktoren minimiert werden sollen (S. 246/247) und führt sechs Aspekte der Konstruktvalidität auf, die bei der Beurteilung aller Messung im Bildungsbereich dienen können (S. 248 ff.). Diese sollen hier aber nicht extra aufgeführt werden.

Vgl. auch Messick 1994

really determines the test tasks is the effect they will have: on student learning, curriculum, educational policy, and so forth. We call this EFFECT-DRIVEN TEST DEVELOPMENT […] the pragmatic nature of the exam is the complex of impact it has upon its users, its testtakers, and the nation as a whole. […] The exam in question becomes what it does: its role is its meaning" (2006, S. 232, Hervorhebung im Original).

Die Idee beschränkt sich natürlich nicht auf Sprachenlernen und Sprachprüfungen, sondern ebenso auf andere Bildungsgebiete. Auch gibt es mittlerweile einiges an Literatur zu diesem Thema, das unter dem Stichwort „Bildungsinnovation" *(educational innovation)* zusammengefasst werden kann. Hier sollen nur ganz knapp die relevanten Hauptpunkte aufgeführt werden, ohne sie weiter zu vertiefen.

Wall stellt die Grundfrage so: „Is there a set of procedures which can be followed to make sure that tests behave ‚the way they should' or is the process of producing ‚washback' unpredictable?" (Wall 1997, S. 334)[15] Das Ziel ist es, möglichst viele Störfaktoren und Gründe, die zum Scheitern führen könnten, vorherzusehen und ihnen vorzubeugen. Der Einführung eines neuen *high-stakes-Tests* sollte eine gründliche Analyse des bestehenden Bildungssystems, also des Kontextes, in dem er eingeführt werden soll, vorausgehen (vgl. Cheng 1997, S. 51). Diese kann dann später auch als *baseline*-Studie für die Überprüfung des Washback-Effekts verwendet werden.

Shohamy et al. nennen zunächst drei Faktoren, die sich auf die Intensität des vermutlichen Washback eines Sprachtests auswirken können, nämlich den Status der Sprache, die getestet wird, die Art des Tests – hier reicht das Spektrum von der Klausur in einer Schulklasse bis zur landesweiten, zentralisierten Prüfung – und den Zweck, wofür die Testergebnisse verwendet werden (vgl. Shohamy/Donitsa-Schmidt/Smadar 1996, S. 299/300). Im Falle des Englischteils der Aufnahmeprüfungen in Japan weisen alle drei Faktoren daraufhin, dass ein Waschback-Effekt von höchster Intensität zu erwarten ist: Der Status des Englischen wird allgemein als sehr viel höher als der anderer Sprachen eingestuft. Auch im Vergleich zu anderen Fächern kann der Stellenwert als hoch bezeichnet werden. Da es sich zumindest beim *sentâ-shiken* um einen landesweiten, zentralisierten Test handelt und auch die Aufnahmeprüfungen der einzelnen Universitäten eine große Anzahl von Testnehmern haben, kann man von einer ausgesprochen breiten Streuung sprechen. Und schließlich haben die Testergebnisse ausschließlich selektive Funktion und wirken sich nachhaltig auf die Berufliche Laufbahn des Testnehmers aus.

Des Weiteren können drei Ebenen unterschieden werden, auf denen sich der Einfluss eines *high-stakes*-Tests bemerkbar machen kann Lerninhalte, Lehrmethode und die Einstellungen der Beteiligten *(content, methodology and attitude)*. Dabei vollzieht sich der Einfluss auf den Inhalt am direktesten und mehr oder weniger automatisch, während es auf den beiden anderen Ebenen sehr viel schwieriger ist, positiven Washback zu erzeugen oder nachzuweisen (vgl. u. a. Alderson/Banerjee 2001 S. 214). Popham (1987, S. 680) nennt in Hinblick auf den Inhalt drei Bedingungen, die für positiven Washback notwendig sind: Validität, für die Lerner relevante Lernziele *(defensible content)* und eine übersichtliche Anzahl von Lernzielen *(managable number of targets)*, um es zu ermöglichen, diese an andere Beteiligte, vor allem

[15] Vgl. auch Wall 2000, wo sie einen Überblick über die Literatur zum Thema gibt und wichtige Faktoren sowie 12 Grundthesen der Theorie über Bildungsinnovation aufführt.

Lehrende, zu vermitteln. Inhaltliche Aspekte können hier nicht eingehender diskutiert werden. Es sei darauf erwiesen, dass der GER hier eine große Hilfe sein kann.

Es gibt zwei Gründe, weshalb der Einfluss von Tests nur schwer kontrollierbar und auch schwer nachweisbar ist. Zum einen gibt es eine größere Anzahl von Beteiligten, die sich jeweils gegenseitig beeinflussen, zum anderen sind es nicht nur objektive Tatsachen, die bei diesem Prozess eine Rolle spielen, sondern die subjektiven Vorstellung der Beteiligten, bzw. ihr Verständnis von Neuerungen, wie etwa des Tests, der eingeführt werden soll. "What may count is not the objective difficulty of the test, but the test-takers' subjective perception of its difficulty that may potentially cause Washback" (Watanabe 2001, S. 108). Es müssen also nicht nur objektiv existierende Faktoren in die Rechnung mit einbezogen werden, sondern auch antizipiert und untersucht werden, wie diese von Beteiligten verstanden und aufgenommen werden. Als Beteiligte wären in diesem Zusammenhang Testanbieter, Lehrerausbilder, Lehrer, Lerner, Eltern, für Curricula Zuständige, Bildungspolitiker, Administratoren zu nennen. Kennedy (1988 nach Wall 1996, S. 342) gibt eine Hierarchie der beteiligten Gruppen (*hierarchy of inter-relating subsystems*) bei Innovationen im Bildungswesen an: *cultural – political – administrative – educational – institutional – classroom*. Auch der GER äußert sich in Kap. 6.3 zu den Beteiligten. Der Erfolg bei der Einführung eines neuen Tests in ein Bildungssystem wird also weitgehend von der Kommunikation zwischen den Beteiligten abhängen, vor allem aber davon, wie gut es den Testanbietern gelingt, die Inhalte und Vorteile eines handlungsorientierten Tests zu kommunizieren.

Einer der wichtigsten Punkte, vor allem auch in Hinblick auf Japan, ist dabei mit Sicherheit die Lehrerausbildung (vgl. u. a. Heyneman/Ransom 1990; Wall 1997, Popham/Kellaghan/Greaney 1992, Schocker-von Ditfurth 2003, Watanabe 2004b). Testinhalte und Lernziele müssen den Lehrern vermittelt werden (vgl. Popham 1987, S. 680/681). Dies gilt vor allem auch für didaktische Ansätze, bei denen meiner Meinung nach noch sehr viel Nachholbedarf besteht. Kommunikativer Unterricht hat sich in Japan sowohl auf schulischer wie auf universitärer Ebene bisher noch erstaunlich wenig durchgesetzt. Dies mag an tradierten Lerner- und Lehrerrollen, unzureichender Lehrerausbildung, Lehr- und Lerntraditionen und der häufigen Fixierung auf Sprachwissenstests liegen. Insofern wird auch einiges an Druck notwendig sein, um das Unterrichtsverhalten von Lehrenden zu verändern und sie dazu zu bewegen, auf ein breiteres didaktisches Spektrum zurückzugreifen, dass es ihnen ermöglicht, handlungsorientierten Unterricht zu halten.[16] Es scheint eine Art Konsens zu sein, dass die Lehrer das problematischste Glied in der Kette sind. Genau da aber könnte ein Test die Möglichkeit darstellen, Veränderung zu bewirken, da Lehrer, vor allem an japanischen Schulen, immer unter Druck stehen, die Schüler auf die Aufnahmeprüfungen vorzubereiten. Zum einen könnte ein handlungsorientierter *high-stakes*-Test für diejenigen Lehrer, die fähig und willens sind, kommunikativen Unterricht zu machen, die Bahn ebnen, indem er die Restriktionen eines *high-stakes*-Tests, der vorwiegend Sprachwissen prüft, beseitigt, und indem er diejenigen Lehrer, die

[16] Dies veranschaulicht ein Beispiel aus Israel, in dem Shohamy/Donitsa-Schmidt/Smadar (1996, S. 314) im Vergleich zwischen einem Arabischtest und einem Englischtest zeigen, dass ein gewisser Druck notwenidg ist, um das Lehrerverhalten zu verändern: „Current interviews held with teachers have shown that once teachers learnt that the result had no personal immediate effect on them, they became relaxed and fearless and thus the effect of the test decreased".

Veränderungen bisher verweigert haben, weil sie sich nicht für Fremdsprachendidaktik interessieren, dazu zwingen könnte, sich weiterzubilden, wenn sie in ihrem Beruf überleben wollen. Auch hierzu äußert sich der GER: „Von Lehrenden wird erwartet, dass sie die Fortschritte der Schüler / Studierenden kontrollieren und Wege finden, Probleme, die beim Lernen auftauchen, zu erkennen, zu analysieren und zu beheben; außerdem ist es Aufgabe von Lehrenden, die individuellen Lernfähigkeiten ihrer Schüler / Studierenden weiterzuentwickeln. Es ist erforderlich, dass Lehrende die Vielfältigkeit der Lernprozesse verstehen." (Kap. 6.3.4).

Wie Cheng (1997) am Beispiel Hong Kongs zeigt, beteiligten sich neben den Testanbietern auch die Lehrwerksverlage durch Weiterbildungsveranstaltungen für Lehrende aktiv daran, die Lernziele und das dazu notwendige didaktische Rüstzeug zu vermitteln. Gogolin (2003, S. 91) gibt in Hinblick auf den GER allerdings zu bedenken: „Das Instrumentarium Gemeinsamer Europäischer Referenzrahmen setzt also nicht unmittelbar am Können und den tiefsitzenden Berufsroutinen von Lehrkräften an" (S. 93). Sie betont dass der Einarbeitungsprozess von Lehrenden beobachtet werden muss und es notwendig ist zu untersuchen, „welche Leistungen ihnen abverlangt werden, welche davon sie ohne weiteres erbringen können und welche nicht" (ebd.).[17] Lehrerausbildung sollte überlegt und ihrerseits didaktisch fundiert sein. Sie sollte auch anschaulich den geforderten Unterrichtsstil selbst umsetzen, die Lehrenden also zur Reflektion ihres Lehrhintergrundes anregen und möglichst projektorientiert sein: Innovation muss anschaulich erfahrbar sein, sonst bleibt keine nachhaltige Wirkung auf Neuorientierung und Reflexion des Lehrerverhaltens und Selbstverständnisses als Lehrer. Lehrer müssen „Prinzipien kooperativer Arbeitsformen kennen […]. Diese berufsfeldbezogene Vorbereitung können rein wissensvermittelnde Seminare nicht leisten." (Schocker-von Ditfurth 2003, S. 168). Dasselbe ist natürlich auch für die Ausbildung von Lehrpersonal, nicht nur bei der Weiterbildung, wünschenswert. Auch der GER kann bei der Lehrerausbildung und Weiterbildung von Hilfe sein, indem er mit seinen Fragestellungen zur Reflektion über den Unterricht anregt. „Der Gemeinsame europäische *Referenzrahmen* will helfen die Barrieren zu überwinden, die aus den Unterschieden zwischen den Bildungssystemen in Europa entstehen und die der Kommunikation unter Personen, die mit der Vermittlung moderner Sprachen befasst sind, im Wege stehen. Er stellt Werkzeuge zur Verfügung für Verantwortliche im Bildungswesen, für Lehrwerkautoren, Lehrende, Lehrerausbilder, Prüfungsanbieter usw., die ihre Tätigkeiten reflektieren wollen, um ihre Bemühungen einzuordnen und zu koordinieren sowie sicherzustellen, dass sie die tatsächlichen Bedürfnisse der Lernenden, für die sie verantwortlich sind, befriedigen." (GER, Kap. 1.1)

In Bezug auf Japan wären tiefer gehende Studien darüber notwendig, wie es genau um die didaktische Ausbildung von Lehrpersonal, das an verschiedenen Bildungseinrichtungen tätig ist, aussieht, wie motiviert und kompetent die Lehrenden sind, ihren Unterricht handlungsorientierter auszurichten,[18] zu welchen

[17] Einen vagen Einblick in die Voraussetzungen von Lehrern gibt Watanabe (1996) indem er den Unterricht von zwei Lehrern an einer *yobikô* untersucht und wie verschiedene Unterrichtsziele ihren Unterricht verändern.

[18] Watanabe gibt einige Statements von Lehrern, die illustrieren, dass sie sehr beschränkte didaktisch-methodische Kenntnisse haben (degree of teachers' familiarity with a range of teaching methods). Z.T. sind sich die Lehrer dessen auch selbst bewusst. Ein Lehrer z.B. gibt zu, dass er nicht wisse, wie er Hörverstehen unterrichten sollte, ein anderer sagt, dass

Teilen die gegenwärtige Ausbildungssituation auf Lerntraditionen, Prüfungen oder andere Faktoren zurückgeführt werden kann und wie die Einstellungen Lehrender an verschiedenen Bildungseinrichtungen und verschiedener Altersklassen gegenüber einer Veränderung der Unterrichtspraxis aussehen.

Watanabe (2004b, S. 140) betont, dass die Lehrer ein verzerrtes Bild der Testinhalte der Aufnahmeprüfungen haben, das sich auch auf ihren Unterricht und ihre Einstellung auswirkt. Dies liegt zum einen an der unübersichtlichen Vielzahl der Aufnahmeprüfungen der verschiedenen Universitäten, die in der Regel auch nicht von professionellen Testanbietern, sondern von Lehrpersonal mit ganz anderem akademischen Hintergrund erstellt werden, und zum anderen an der mangelnden Augenscheinvalidität der Prüfungen, deren Inhalte von den Anbietern, also den Universitäten, auch nicht an andere Beteiligte, vor allem Lehrer, kommuniziert werden.

Ein wichtiger Aspekt ist also die Transparenz in Hinblick auf Lernziel und Validität, die sich, wie Popham hervorhebt, auch in der konzeptionellen Nähe der Beschreibung von Test- und Lernzielen in Hinblick auf die Curriculumbildung niederschlägt. „Those who develop high-stakes tests for measurement-driven instruction must conceptualize the skills or knowledge to be tested in such a way that teachers can use the targeted skills and knowledge to design effective instructional sequences“ (Popham 1987, S. 680). Auch Watanabe (2004, S. 142) fordert in Hinblick auf Japan höhere Augenscheinvalidität um die Inhalte der Aufnahmeprüfungen für die Lehrenden fassbarer zu machen.

Zusammenfassend kann also gesagt werden, dass Lehrerausbildung und ausreichende Kommunikation zwischen Testanbieter (bzw. Innovator), Lehrern und anderen Beteiligten die vermutlich wichtigste Bedingung ist, die erfüllt werden muss. Hierzu ist Transparenz in Hinsicht auf Lerninhalte und die Gütekriterien unabdingbar. Ein neuer Test sollte auch nicht als Endziel betrachtet werden, sondern der Washback des neu eingeführten oder veränderten, handlungsorientierten Tests soll nur der Anstoß sein, um Änderungen hin zu kommunikativem Unterricht und autonomerem Lernverhalten zu initiieren. Dabei muss darauf geachtet werden, wie substanziell oder formal die Neuerungen von Lehrenden und Lernenden aufgenommen werden (vgl. Cheng 1997, S. 52). Die Änderung oder Einführung eines Tests selbst reicht meistens noch nicht, um eine Innovation des Bildungssystems hervorzurufen aber sie ist eine notwendige Voraussetzung dafür. Solange der alte Test bleibt, besteht keine Chance auf Verbesserung, wie Alderson auch in Hinblick auf die Situation in Sri Lanka anmerkt: „Although efforts are being made to improve matters, it is recognized that all will be in vain if the exams do not change“ (Alderson 1986, S. 104). Große, zentralisierte Tests wie z.B. die Aufnahmeprüfungen in Japan durch vollkommen handlungsorientierte Tests wie sie etwa Messick (1996) charakterisiert, abzulösen, ist wohl vom logistischen Aufwand her nicht möglich. Es sollte allerdings graduell eine deutliche Verbesserung möglich sein. Auf jeden Fall muss ein Kompromiss gefunden werden.

---

es schwierig sei, andere Unterrichtsmethoden zu verwenden, als die mit denen er früher unterrichtet worden sei. (Watanabe 2004b, S. 140)

2. Wie kann der GER von Nutzen sein?

Da es auch auf Japanisch bereits einiges an einführender Literatur zum GER[19] sowie eine Übersetzung des GER ins Japanische[20] gibt, sollen hier nur die für das oben dargelegte Thema relevanten Charakteristika des GER knapp aufgeführt werden.

„The direct instruction model under the influence of behaviorism – tell-show-do approach – does not match how students learn, nor does it take into account students' intentions, interests, and choices. Teaching that fits the cognitive-constructivist view of learning is likely to be holistic, integrated, project-oriented, long-term, discovery-based, and social. Likewise, testing should aim to be all of these things too" (Cheng/Curtis 2004, S. 15). Genau hier setzt auch der GER an, und es ist der Grund warum er als Werkzeug für eine Erneuerung der Testlandschaft in Japan von Nutzen sein kann. Im Folgenden sollen hierfür einige Argumente aufgeführt werden.

Anschließend an das oben gesagte sei zunächst darauf hingewiesen, dass es das erklärte Ziel des GER ist, ein Werkzeug zur Kommunikation und Transparenz zu liefern. Wie in Kap. 1.3 aufgeführt, sind es seine Ziele, „die Kooperation zwischen den Bildungseinrichtungen in den verschiedenen Ländern zu fördern und zu erleichtern, die gegenseitige Anerkennung der sprachlichen Qualifikationen auf eine solide Basis zu stellen, Lernende und Lehrende, Autoren von Sprachkursen, Prüfungsanbieter und die Bildungsverwaltung dabei zu unterstützen, ihre Bemühungen in diesen Rahmen einzubetten und sie zu koordinieren." Dies geschieht durch differenzierte und erprobte Kannbeschreibungen und Niveaustufen, wobei immer die Bedürfnisse der Lernenden im Mittelpunkt stehen. Dies ist wohl auch das wichtigste Argument, weshalb man mit Hilfe des GER versuchen sollte, ein Gegengewicht zu den Bedürfnissen der Testanbieter von Sprachwissenstests, die vornehmlich dem Zwecke der Selektion dienen, zu erhalten, sodass ein Kompromiss gefunden wird, der logistisch machbar ist, inhaltlich und formal aber trotzdem weitgehend die Bedürfnisse Lerner, die den Test ablegen müssen, berücksichtigt, indem sinnvolle und realistische Lernziele formuliert und möglichst genau beschrieben werden (vgl. Kap. 2.2).

2.1. Handlungsorientierung

Little kommt in seinem state-of-the-art-Aufsatz in „Language Teaching" zu dem Schluss, „[…] there is no doubt that the CEFR's action-oriented approach can be drawn on to support the development of communicative language tests in contexts where they may still be a novelty" (2006, S. 185). „Action-oriented", also „handlungsorientiert" bedeutet, dass die Fähigkeiten, die sich Lernende aneignen müssen, wenn sie eine oder mehrere Sprachen lernen, umfassend beschrieben und beim Lehren, Lernen und Prüfen berücksichtigt werden. Dabei ist deklaratives Sprachwissen nur einer von vielen Aspekten. „Der

---

[19] Siehe u. a. Fujiwara 2004 oder 『ヨーロッパにおける日本語教育事情とCommon European Framework of Reference for Languages』 von der Japan Foundation (kann unter http://www.jpf.go.jp/j/japan_j/publish/euro/ vollständig runtergeladen werden).

Einen knappen aber relativ umfassenden und aktuellen Überblick auf Englisch gibt Little 2006.

[20] Als Buch, übersetzt von Yoshijima Shigeru, Ôhashi Rie u. a. 2004 im Asahi-Verlag erschienen unter dem Titel 『外国語の学習のためのヨーロッパ共通参照枠』 *(Gaikokugo no gakushû no tame no yôroppa kyôtsû sanshô-waku)*. Vor allem Kap. 2 bietet einen Überblick über die Grundideen des GER.

hier gewählte Ansatz ist im Großen und Ganzen ‚handlungsorientiert', weil er Sprachverwendende und Sprachenlernende vor allem als ‚sozial Handelnde' betrachtet, d.h. als Mitglieder einer Gesellschaft, die unter bestimmten Umständen und in spezifischen Umgebungen und Handlungsfeldern kommunikative Aufgaben bewältigen müssen, und zwar nicht nur sprachliche" (GER, Kap. 2.1). Genau dies ist der Ansatzpunkt, von dem aus Prüfungen und dadurch schließlich der Fremdsprachenunterricht in Japan verbessert werden können, indem Kontextbezug (Domänen), prozeduralen Fertigkeiten (Strategien), Lernerautonomie (Entwicklung kognitiver und metakognitiver Fähigkeiten), Selbstevaluation, lernpsychologischen Aspekten usw. der Stellenwert eingeräumt wird, der ihnen zukommt.

Domänen sind Situationen, die den Kontext für den Sprachgebrauch darstellen. „Jede Sprachverwendung findet im Kontext einer bestimmten Situation innerhalb eines der *Lebensbereiche (Domänen)* statt, in denen das soziale Leben organisiert ist" (GER, Kap. 4.1.1). Als Beispiel für die Ausarbeitung von Redemitteln für verschiedene Domänen kann für die Deutsche Sprache etwa „Profile Deutsch" (Glaboniat/Müller 2005) angeführt werden. Auch für Englischprüfungen in Japan sollte es nicht zu schwierig sein, anhand des offen gehaltenen GER und konkreten Beispielen wie etwa aus „Profile Deutsch" und verschiedenen Lehrwerken, die mit dem GER kompatibel sind, geeignete Domänen und die entsprechenden Redemittel festzulegen. Jedenfalls sollte man sich die „pragmatisch-funktionale Sicht auf Sprache", die dem GER zugrunde liegt (Krumm 2003, S. 121) zu Nutzen machen, um in der Unterrichtspraxis endlich eine Verschiebung vom Primat der Vermittlung formalsprachlichen Wissens zu inhaltlicher, strategischer und kommunikativer Kompetenz zu vollziehen. „Der Schwerpunkt einer kommunikativen Aufgabe liegt auf ihrer erfolgreichen Bewältigung und im Mittelpunkt steht folglich die inhaltliche Ebene, während Lernende ihre kommunikativen Absichten realisieren." (GER, Kap. 7.1) „Kommunikative didaktische Aufgaben (im Gegensatz zu Übungen, bei denen das dekontextualisierte Einüben von Formen im Mittelpunkt steht) haben das Ziel, die Lernenden aktiv an sinnvoller Kommunikation zu beteiligen" (ebd.). Dies bezieht sich selbstverständlich nicht nur auf den Bereich der mündlichen Kommunikation. Auch beim Lesen z.B. müssen die Lernenden mit verschiedenen Lesestilen (kursiv, selektiv, detailliert usw.) die an bestimmten konkreten Aufgaben und kommunikativen Kontexten festgemacht sind, sowie mit Lesestrategien (Rezeptionsstrategien wie die Anwendung von Weltwissen, Inferieren, Hypothesen aufstellen und testen usw.), der selbständigen Benutzung von Hilfsmitteln (Wörterbüchern, Grammatiken usw.) vertraut gemacht werden, sonst besteht, wie beim GÜM-Unterricht, die Gefahr, dass Lernende, wie es in Japan häufig zu beobachten ist, nur noch detailliert lesen, was ihnen extensives Lesen unmöglich macht und zu Frustration führt.

Der holistische Ansatz des GER lässt sich auch in der folgenden Definition erkennen: „Kompetenzen sind die Summe des (deklarativen) Wissens, der (prozeduralen) Fertigkeiten und der persönlichkeitsbezogenen Kompetenzen und allgemeinen kognitiven Fähigkeiten, die es einem Menschen erlauben, Handlungen auszuführen." (GER, Kap. 2.1, vgl. auch Kap. 5 über Kompetenzen). Dabei werden Strategien als „Gelenkstellen zwischen den Ressourcen der Lernenden (Kompetenzen) und dem, was sie mit ihnen tun können (kommunikative Aktivitäten) betrachtet. In den Abschnitten in Kapitel 4, die sich mit Interaktions- und mit Produktionsstrategien befassen, werden folgende Strategiegruppen beschrieben: (a) Planung von

Handlungen, (b) Ressourcen sinnvoll gewichtet einsetzen und Defizite bei der Ausführung von Aktivitäten kompensieren, (c) Kontrolle (monitoring) der Ergebnisse und – wenn nötig – Reparaturhandlungen“ (GER, Kap. 3.4). Dies sind auch die Parameter, nach denen die Skalierungen mit Kann-Beschreibungen in Kap. 4.4 aufgebaut sind. Insofern spielen sie eine zentrale Rolle bei der Umsetzung, dessen, was man gelernt hat, und dürfen weder in Tests noch im Unterricht unberücksichtigt bleiben, sei es, dass sie explizit geübt werden oder wenigstens durch kommunikative, interaktive Aufgaben zur Anwendung kommen. Hierzu befindet sich in Kap. 4.4 des GER ein reichhaltiger Fundus an systematisch geordneten Skalierungen zu Kommunikations- und Kompensationsstrategien: sprachproduktiv, -rezeptiv und interaktiv (z.B. auch Lese- und Hörverstehen), Rezeptionsstrategien, sowie Auflistungen von Arten und Strategien mündlicher Kommunikation (Interview, Verhandlung, Diskussion …) sowie schriftlicher (Korrespondenz, Verträge, Berichte …), Interaktionsstrategien (Sprecherwechsel, kooperieren, um Erklärung bitten, Missverständnisse aufklären …), Arten und Situationen beim Dolmetschen und Übersetzen, Strategien der Sprachmittlung (Vorbereiten, Antizipieren, Hilfsmittel verwenden …) und sogar paralinguistische Mittel (Körpersprache und prosodische Mittel). Da sie bei der Sprachverwendung tatsächlich zum Einsatz kommen, sollten sie auch Gegenstand von Unterricht und Prüfungen sein.

2.2. Kann-Beschreibungen und Niveaustufen

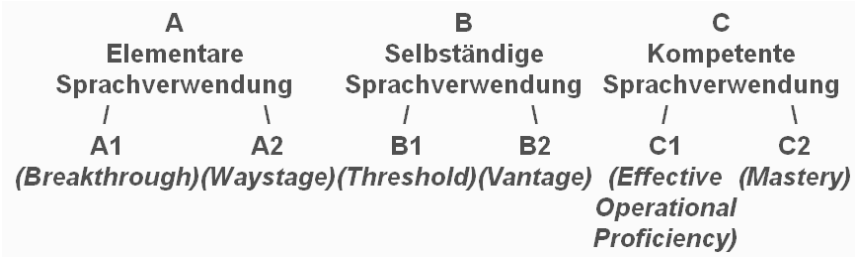
Die praktischen Werkzeuge des GER sind seine Kann-Beschreibungen. Sie gehen auf die Dissertation von Brian North (2000) und die *can-do-statements* von ALTE (Association of Language Testers in Europe) zurück, die ein fünfstufiges Skalensystem hatte. Sie sind durch viele Jahrgänge von Tests erprobt und validiert (vgl. Quetz 2003, S. 145 ff. und GER, Kap. 3.1 und Anhang A). Durch die Deskriptoren sollen positive Beschreibungen etwa in Hinblick auf die verschiedenen Teilfertigkeiten geliefert werden, was ein Lerner bereits kann. Dabei werden auch qualitative Deskriptoren verwendet (Korrektheit, Flüssigkeit usw.). Es würde hier zu weit führen, den Aufbau der Kann-Beschreibungen weiter auszuführen. Es wird empfohlen, sich die Kann-Beschreibungen im GER einfach selbst anzuschauen, z.B. in Kap. 3.3. Zur Entwicklung der Deskriptoren siehe Anhang A des GER.

Der Vorteil der Kann-Beschreibungen liegt darin, dass sie sich für die Formulierung von Lernzielen ebenso wie zur Curriculum- und Unterrichtsplanung, zur Gestaltung von Tests und etwa anhand des Europäischen Sprachenportofolios[21] zur Selbstevaluation eignen. Sie lassen sich auch zum Erstellen von Gruppenprofilen verwenden, um Curricula oder Prüfungen zu speziellen Lernzielen und für bestimmte Lernergruppen zu gestalten.

Der Referenzrahmen teilt die Sprachkompetenz in sechs Niveaustufen ein[22]:

[21] http://www.coe.int/T/DG4/Portfolio/

[22] Zur Beschreibung der Referenzniveaus siehe GER Kap. 2.2 und 3. Eine tabellarische Übersicht befindet sich in Kap. 3.3.



(Diagramm aus GER, Kap. 3.1)

Die Niveaustufen mit ihren Deskriptoren[23] haben zwei Funktionen. Zum einen sollen sie für Transparenz sorgen, indem sie ein Bezugssystem bieten, mit dessen Hilfe man Vergleichbare Aussagen über Testergebnisse erhalten kann (vgl. Glaboniat/Müller 2006, S. 15). „Indem er eine gemeinsame Basis für die explizite Beschreibung von Zielen, Inhalten und Methoden zur Verfügung stellt, erhöht der *Referenzrahmen* die Transparenz von Kursen, Lehrplänen und Richtlinien und von Qualifikationsnachweisen" (GER, Kap. 1.1). Es versteht sich, dass genau diese Funktion bei der Einführung eines neuen Tests in ein System unabdingbar ist, um die Kommunikation zwischen den verschiedenen Beteiligten und die Augenscheinvalidität des Tests zu gewährleisten.

Die Zweite Funktion ist es, Kompatibilität zwischen Lehrwerken, Unterricht und Sprachprüfungen zu schaffen. Prüfungen und Niveaus an verschiedenen Institutionen können nun eingeschätzt und aufeinander abgestimmt werden. Die Kann-Beschreibungen helfen, Lern- und Prüfungsziele zu vereinheitlichen. Dass dies besonders für Innovation durch Washback nützlich ist, liegt auf der Hand. Krumm weist auf die Gefahr hin, dass Unterrichtsinhalte beschnitten werden könnten. Dadurch „droht ein Backwash-Effekt derart, dass nur noch das in den Lehrplänen und Lehrwerken Platz findet, was in diesem Sinne durch die Kann-Bestimmungen fixiert und abprüfbar wird" (2003, S. 124). Krumm kritisiert, dass dies schon jetzt bei der Lehrwerksproduktion an der Anpassung der Lehrwerke an die Niveaustufen und Zertifizierung durch die Prüfungsinstanzen ablesbar ist. Er sieht vor allem „Profile deutsch" als Gefahr. (ebd.). Dabei bezieht er sich auf Lehrwerke für Deutsch als Fremdsprache. Hierzu kann man zwei Einwände ins Feld führen. Zum einen kann durchaus gesagt werden, dass die meisten Deutsch-Lehrwerke, die momentan auf dem Markt sind und sich an den Niveaustufen des GER orientieren ein sehr hohes inhaltliches und didaktisches Niveau haben und weitgehend dem aktuellen Stand der Sprachlehrforschung gerecht werden. Zum anderen soll im Hinblick auf Japan genau das erreichen, was Krumm kritisiert, nämlich dass sich Lehrwerke und Unterrichtspraxis an die Standards des GER anpassen, indem sie die Grundideen des GER aufnehmen.

2.3. Unterricht, Curriculum und Lernen

„Eins der vorrangigen Ziele des *Referenzrahmens* ist es, die verschiedenen am Sprachenlernen und -lehren Beteiligten zu ermutigen und zu befähigen, die Anderen so klar wie möglich über ihre Ziele und Absichten

[23] In Kapitel 3.8 werden drei Arten von Sprachkompetenzskalen nach ihrer Funktion unterschieden und beschrieben: „(a) Skalen für Benutzer (**benutzerorientierte Skalen**), (b) Skalen für Beurteilende (**beurteilungsorientierte Skalen**) und (c) Skalen für Testautoren (**aufgabenorientierte Skalen**)"

zu informieren. Ebenso wichtig ist es aber, dass sie auch über die Methoden, die sie benutzen, und über die Resultate, die sie erzielen, informieren können“ (GER, Kap. 2.3.1). Es kann wohl gesagt werden, dass diese Transparenz, zumal in Japan, in den meisten Lehrveranstaltungen nicht zu finden ist. Viele Lehrende und Lernende scheinen keine genaue Vorstellung von den Lernzielen und didaktischen Ansätze zu haben, mit denen sie diese erreichen wollen. Ein weiterer für die Situation in Japan höchst relevanter Punkt wird im GER explizit angesprochen: Der GER steckt „Parameter, Kategorien, Kriterien und Skalen ab, auf die die Benutzer zurückgreifen können und die sie vielleicht dazu anregen, ein größeres Spektrum von Optionen in Betracht zu ziehen als vorher - oder sogar zuvor ungeprüfte Annahmen über das Lernen und Lehren von Sprachen in Frage zu stellen, die in ihrem Arbeitskontext als Tradition gelten“ (Kap. 2.3.2). Als ein Beispiel könnte hier etwa die Vermittlung von Grammatik genannt werden. Grammatikvermittlung die mit den Grundsätzen des GER übereinstimmen würde, und die sich auch in allen modernen DaF-Lehrwerken, die sich auf den GER beziehen, widerspiegelt, ist induktiv, d.h. der Lerner wird dazu angeleitet, sich grammatische Regeln selbst zu erschließen, und sie bevorzugt eine didaktische Grammatikprogression, bei der einfachere, häufig gebrauchte grammatische Strukturen zuerst vorkommen. Außerdem ist die Grammatikprogression nicht zu steil, sodass die Lernenden die Strukturen auch verinnerlichen können. In der Unterrichtspraxis allerdings wird häufig versucht linguistisch-systematische Grammatik bei sehr steiler Progression durch Lehrermonolog zu vermitteln. Mithilfe des GER formulierte Lernziele, das Verständnis seiner didaktischen Grundlagen, sowie das Studium bereits existierender Lehrwerke und Prüfungen, die auf dem GER basieren, könnten hier durchaus eine große Hilfe sein. Jedenfalls regt der GER explizit dazu an, sich über didaktische Aspekte des Unterrichts Gedanken zu machen. Dabei gibt er auch konkrete Anleitungen für Lehrende (z.B. zum Thema Fehlerkorrektur in Kap. 6.5) und es ist ein ganzes Kapitel dem Thema kommunikative Aufgaben gewidmet (Kap. 7).[24]

Ein wichtiger Pluspunkt des GER ist, dass er dazu auffordert und dabei hilft, die Verbindung zwischen erklärtem Ziel des Curriculums (und auch des Tests), seinen Teilen und der konkreten, didaktisch-methodischen Umsetzung im Unterricht zu überprüfen. Dies könnte helfen, der Gefahr vorzubeugen, dass selbst trotz einer Veränderung der Lernziele, des Tests und der Materialien die Unterrichtspraxis nicht tiefgreifend berührt wird. Diese Veränderung des Verhaltens und der Einstellungen der Lehrenden wurde ja oben bereits als großes Problem bei jeder Innovation dargestellt. Interessant ist hier z.B. der Fragenkatalog zur Curriculumgestaltung in Kap. 8.4.3, der zeigt, wie stark sich der GER an den Bedürfnissen der Lernenden orientiert (lernerzentrierter Ansatz).

„Hinzuzufügen ist, dass man in allen Fällen im Unterricht aller Sprachen an irgendeinem Punkt Zeit dafür reservieren sollte, über die Methoden, mit denen die Lernenden unterrichtet werden, und die Lernwege, für die sie sich entschieden haben, nachzudenken. Das bedeutet, dass in jedem schulischen Curriculum Freiräume vorgesehen sein müssen für die allmähliche Entwicklung eines ‚Lernbewusstseins’ und für die Einführung einer allgemeinen Spracherziehung, die es den Lernenden erlauben, metakognitive Kontrolle in Bezug auf ihre eigenen Kompetenzen und Strategien zu entwickeln“ (GER, Kap. 8.3.2 unten). Nicht zuletzt wegen des explizit lernerzentrierten Ansatzes, der dem GER durchweg zugrunde liegt, wäre

[24] Kapitel 8 des GER ist dem Thema „Curriculum“ gewidmet.

es wünschenswert, wenn er in Japan mehr Beachtung und konkrete Umsetzung erfahren würde. So sollen den Lernenden nicht nur Sprachwissen und sprachliche Inhalte vermittelt werden, sondern Allgemeine Kompetenzen (Kap. 5.1) wie deklaratives Wissen *(savoir),* zu dem auch Weltwissen, Soziokulturelles Wissen und interkulturelles Bewusstsein zählen, prozedurales Wissen *(savoir faire),* persönlichkeitsbezogene Kompetenz *(savoir être),* also Einstellungen, Motivation, Überzeugungen, kognitiver Stil und Persönlichkeitsfaktoren, sowie Lernfähigkeit *(savoir apprendre)* wie Lernstrategien, heuristische Fähigkeiten, sowie kommunikative Sprachkompetenzen (Kap. 5.2.1) wie linguistische (lexikalische, grammatische usw.), soziolinguistische (Sprecherwechsel, Höflichkeitskonventionen, sprachliche Register usw.) und pragmatische Kompetenzen (Diskurskompetenz, fragen, beschreiben, gemeinsam organisieren usw.). Die meisten dieser Aspekte treten im „traditionellen Unterricht" weit in den Hintergrund und werden nicht ausreichend vermittelt und geübt. Immerhin finden sie in modernen Lehrwerken, die mit dem GER kompatibel sind, weitgehend Beachtung.

Abschließen sei noch darauf hingewiesen, dass es u. a. mit dem Europäischen Sprachenportofolio Instrumente zur Selbstevaluation gibt, die aus dem GER hervorgegangen sind und mit Deskriptoren in der Ich-Form arbeiten. Diese können durchaus zur Transparenz aus der Sicht der Lerner beitragen.

2.4. Sprachprüfungen

„As we have seen, its action-oriented approach entails that it is possible to use the same ‚can do' descriptor to identify a learning target, shape the learning/teaching process, and guide the assessment of learning outcomes. This is perhaps the CEFR's single most innovative feature: that it brings curriculum, teaching/learning and assessment into much closer interdependence than has usually been the case" (Little 2006, S. 187). Dieses Argument spricht wohl für sich selbst und zeigt, weshalb der GER für die Gestaltung von Prüfungen ein wichtiges Instrument darstellen kann. Ein weiterer Punkt, der sich vor allem in Hinblick auf konkrete Fertigkeiten und die Motivation der Lernenden auswirkt, ist die Tatsache, dass der GER mit positiven Kann-Beschreibungen arbeitet, durch die Lernziele und Testinhalte festgelegt und miteinander verbunden werden können. „Die Aufmerksamkeit von Lehrkräften bei der Prüfung von Schülerleistungen richtet sich auf das Nichtvorhandene, Nichtgekonnte oder falsch Gemachte. Diese ‚Fehler-Defizit-Perspektive' eignet sich nicht zur Evaluation aus der man Hinweise auf die Gestaltung von Lehr- und Lernprozessen ziehen kann. Problem ist, dass Lehrer kein Repertoire an Sprachbeschreibungsmitteln hatten" (Gogolin 2003, S. 90, vgl. auch Neuner 2003, S. 142)"

Dabei wird im GER darauf hingewiesen, dass Strategien und prozedurales Wissen durchaus getestet werden können. „Ein Fortschritt im Sprachenlernen zeigt sich am deutlichsten darin, dass Lernende fähig sind, an beobachtbaren sprachlichen Aktivitäten teilzunehmen und kommunikative Strategien einzusetzen. Daher stellen beide eine gute Basis für die Skalierung der Sprachfähigkeit dar" (GER, Kap. 4.4). Wichtig ist allerdings, dass auch sprachproduktive Teile in den Test aufgenommen werden, da dieser äußerst wichtige Aspekt des Lernens im Unterricht sonst immer vernachlässigt werden wird. Sicherlich ist es mit einem Mehr an logistischem Aufwand und schwierigerer Quantifizierbarkeit verbunden, aber dies kann sich

bei einem akzeptablen Kompromiss durchaus auch in Grenzen halten lassen, wie einige Tests, die mit dem GER kompatibel sind, zeigen.[25] Eine nähere Vorstellung von Tests die auf dem GER basieren und ihren Washback-Effekt kann hier nicht vorgenommen werden. Es sein nur darauf verwiesen, dass es auf diesem Gebiet bereits einiges gibt und sich der GER immer mehr verbreitet. Dies ist in sofern Vorteilhaft, als dass man so aus Erfahrungen anderer lernen kann, wenn man in Japan einen neuen Test einführen, bzw. bestehende verändern oder erweitern will.

Außerdem sei noch auf die Möglichkeit hingewiesen, Prüfungen an professionelle Testanbieter zu delegieren. Werff (2006) gibt einen sehr positiven Bericht über eine Campagne zur Förderung einer zweiten Fremdsprache in der Sekundarstufe I in Italien (Progetto Lingue 2000), in deren Rahmen das Goethe-Institut extra neue Prüfungen für Jugendliche auf den Niveaus A1 und A2 (Fit in Deutsch I und II) entworfen hat, die offensichtlich einen sehr positiven Washback hatten. Das erklärte Lernziel war die Beherrschung der gesprochenen Sprache. „Damit hatten die Erziehungsbehörden den Washback-Effekt bewusst eingeplant. In den Richtlinien des Ministeriums wurde verlangt, dass die gesprochene Sprache im Mittelpunkt des Fremdsprachenunterrichts stehen sollte“ (Werff 2006, S. 41). Außerdem betont Werff: „Handlungsorientierte Unterrichtsformen hielten tatsächlich binnen weniger Jahre Einzug in die Klassenzimmer“ (S. 42) und erwähnt einige weitere positive Effekte, wie etwa Partnerschaften und Schüleraustausch, inhaltliche Projektarbeit, Interesse am Land der Zielsprache, Beobachtung des eigenen Lernverhaltens mit Kann-Beschreibungen in den Lehrbüchern, Motivationsschub durch positive Bewertung der Prüfungen dank ihrer Realitätsnähe und Transparenz.

Zwar sind die Rahmenbedingungen dieser Innovation sicherlich anders als in Japan und die eingeführten Prüfungen längst nicht so *high-stakes* wie etwa die Aufnahmeprüfungen in Japan, doch es ist evident, dass gerade *high-stakes*-Test, die einen breiten und nachhaltigen Washback-Effekt haben, am meisten darauf ausgelegt sein sollten, realitätsnahe Inhalte zu testen, um den negativen Washback einzuschränken und möglichst positiven zu erzeugen. Zu diesem Zwecke sollte man unbedingt auf alle verfügbaren Ressourcen, wie etwa den GER, Berichte über erfolgreiche und erfolglose Innovationen in anderen Ländern und auch auf Fachleute zurückgreifen, die Erfahrung in der Erstellung von Prüfungen haben, die weitgehend handlungsorientiert und kommunikativ ausgelegt sind, wenn auch nur in beratender Funktion.

[25] Hier können etwa Test-DaF (http://www.testdaf.de/) oder das ÖSD (http://www.osd.at/) erwähnt werden, die beide zentralisierte *high-stakes*-Test sind, weil sie über Einbürgerung, bzw. Hochschulzugang entscheiden können. Auch die ESOL-Tests (http://www.cambridgeesol.org/) sind in diesem Zusammenhang sicher einen Blick wert. Alle diese Tests haben auch sprachproduktive Teile, einschließlich Sprechen. Bei Test-DaF handelt es sich allerdings um „CD- oder kassettengesteuertes Format“ und nicht um ein direktes Interview mit einem Prüfer.

Literaturangaben

Englisch- und deutschsprachige Quellen

Alderson, J. Charles; Wall, Dianne. Does washback exist? In: *Applied Linguistics,* 14 (2). 1993. S. 115-129

Alderson, J. Charles; Wall, Dianne. Examining Washback: the Sri Lankan Impact Study. In: *Language Testing,* Vol. 10, Nr. 1. 1993. S. 41-70

Alderson, J. Charles; Banerjee, Jayanti. Language Testing and Assessment. State of the Art Review Part 1. in *Language Teaching*, Vol. 34. Cambridge University Press 2001. S. 213-236

Andrews, Stephen. Washback and Curriculum Innovation. In: Cheng, Liying; Watanabe, Yoshinori; Curtis, Andy. *Washback in Language Testing – Research Contexts and Methods.* Lawrence Erlbaum Associates 2004. S. 37-50

Beer, Rudolf. Standards und Leistungsbeurteilung: Bedeutung und grundlegende Funktion. In: *ide (Informationen zur Deutschdidaktik). Zeitschrift für den Deutschunterricht in Wissenschaft und Schule,* 30. Jahrgang, Heft 4, 2006. S. 52-63

Berwick, Richard; Ross, Steven. Motivation after Matriculation: Are Japanese Learners of English Still Alive after Examination Hell? In: *JALT Journal*, 11. 1989. S. 193-210

Bleyhl, Werner. Die sprachliche Leistungsbeurteilung und die Chance zur Verbesserung des Fremdsprachenunterrichts dank des *backwash*-Effekts. In: Bausch, Richard; Christ Herbert u. a. (Hrsg.). *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion: Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Gunter Narr 2003. S. 36-44

Butzkamm,Wolfgang; Plum, Anke. „Nicht für das Leben, für die Prüfung lernen wir" Der Einfluss zentraler schulischer Abschlusstests auf den Unterricht – Ein Beispiel: Deutschunterricht in England. In: *Fremdsprache Deutsch*, Heft 34/2006 („Kompetenzen testen, prüfen, zertifizieren"). Klett 2006. S. 35-39

Cheng, Liying. How Does Washback Influence Teaching? Implications for Hong Kong. In: *Language and Education.* Vol. 11, Nr. 1. 1997. S. 38-54

Cheng Liying. The Washback Effeckt of a Publick Examination Change on Teachers' Perception Toward Their Classroom Teaching. In: Cheng, Liying; Watanabe, Yoshinori; Curtis, Andy. *Washback in Language Testing – Research Contexts and Methods.* Lawrence Erlbaum Associates 2004. S. 3-17

Cheng, Liying; Curtis, Andy. Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In: Cheng, Liying; Watanabe, Yoshinori; Curtis, Andy. *Washback in Language Testing – Research Contexts and Methods.* Lawrence Erlbaum Associates: 2004. S. 3-17

Cheng, Liying. *Changing language teaching through language testing – A Washback study.* Cambridge University Press (Studies in Language Testing 21) 2005

Degen, Ralph. Japanese Students' Foreign Language Acquisition Learner Autonomy: Actual Situation and Some Suggestions for its Promotion. In: *Kagawa kyôiku kenkyû, dai-ichi-gô,* März 2004, 1-17

Frederiksen, John R.; Collins, Allan. A Systems Approach to Educational Testing. Technical Report No. 2. Center for Technology in Education, New York, NY, ED325484. 1990

http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED325484
auch veröffentlicht in: *Educational Researcher*, Vol. 18, Nr. 9, Dez. 1989. S. 27-32

Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. (GER)
Volltext-Versionen online:
- Eglische Version: Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf
- Deutsche Version: http://www.goethe.de/Z/50/commeuro/i3.htm
- Japanische Version: 外国語の学習のためのヨーロッパ共通参照枠： (Gaikokugo no gakushû no tame no yôroppa kyôtsû sanshô-waku):
http://wwwsoc.nii.ac.jp/jgg/jggla/library/cef_verzeichnis.html

Glaboniat, Manuela. Das Papier nicht wert … Zum Problem schulischer Leistungsmessung und Benotung und neue Chancen durch Qualitäts- und Leistungsstandards. In: *ide (Informationen zur Deutschdidaktik). Zeitschrift für den Deutschunterricht in Wissenschaft und Schule,* 30. Jahrgang, Heft 4, 2006. S. 32-51

Glaboniat, Manuela; Müller, Martin; u. a. *Profile deutsch* (CD-ROM Version 2.0 mit Begleitbuch). Langenscheid 2005

Glaboniat, Manuela; Müller, Martin. Note „Sehr gut!" - Aber in Bezug worauf? Referenzrahmen und Profile Deutsch in ihren Auswirkungen auf Prüfungen und Tests. In: *Fremdsprache Deutsch*, Heft 34/2006 („Kompetenzen testen, prüfen, zertifizieren"). Klett 2006. S. 14-21.

Gogolin, Ingrid. Der Gemeinsame europäische Referenzrahmen. In: Bausch, Richard; Christ Herbert u. a. (Hrsg.). Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion: Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts. Tübingen: Gunter Narr 2003. S. 85-94

Grotjahn, Rüdiger. Testtheorie: Grundzüge und Anwendung in der Praxis. In.: Wolff, Armin; Tänzer, Harald. Sprache - Kultur - Politik. Beiträge der 27. Jahrestagung Deutsch als Fremdsprache vom 3. - 5. Juni 1999 an der Universität Regensburg. Universität Regensburg: Fachverband Deutsch als Fremdsprache (Materialien Deutsch als Fremdsprache Bd. 53) 2000. S. 304-341

Heyneman, S. P. und Ransom, A. W. Using examination and testing to improve educational quality. In: *Educational Policy.* 1990. 177-192

Kikuchi, Keita: Revisiting English Entrance Examinations at Japanese Universities after a Decade. In: *JALT Journal* Vol. 28, No. 1, May 2006

Komárek, Friderike. *Der Referenzrahmen und seine praktischen Auswirkungen im Fremdsprachenunterricht.* 2006.
http://www.euregio-egrensis.de/sprachoffensive/kursleitertreffen.php

Krumm, Hans-Jürgen. Der Gemeinsame europäische Referenzrahmen – ein Kuckucksei für den Fremdsprachenunterricht? In: Bausch, Richard; Christ Herbert u. a. (Hrsg.). *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion: Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Gunter Narr 2003. S. 120-126

Krumm, Hans-Jürgen. Müssen jetzt alle dasselbe können? Vor- und Nachteile der Globalisierungsprozesse im Sprachunterricht. In: *Fremdsprache Deutsch*, Heft 34/2006 („Kompetenzen testen, prüfen, zertifizieren"). Klett 2006. S. 30-33

Little, David. The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. In: *Language Teaching,* Vol. 39 (3). Cambridge University Press: 2006. S. 167-190

Madaus, G.F. The influence of testing on the curriculum. In: Tanner, L.N. (Hg.). *Critical Issues in Curriculum: Eighty-seventh Yearbook of the National Society for the Study of Education.* University of Chicago Press: 1988. S. 83–121

Messick, Samuel. The interplay of Evidence and Consequences in the Validation of Performance Assessments. In: *Educational Researcher*, Vol. 23, No. 2. 1994. S. 13-23

Messick, Samuel. Validity and washback in language testing. In: *Language testing,* Vol. 13 (3). 1996. S. 241-256

North, Brian. The development of a common framework scale of language proficiency. New York: P. Lang 2000

Pearson, Ian. Tests as Levers for Change (or ‚Putting First Things First'). In: Chamberlain, Dick; Baumgardner, Robert. *ESP in the classroom: Practice and evaluation,* Vol 128. London: Modern English Publications 1988. S. 98-107

Popham, J. The merits of measurement-driven instruction. In: *Phi Delta Kappan.* 1987. Vol. 68, May, 679–68

Quetz, Jürgen. Der gemeinsame Europäische Referenzrahmen: Ein Schatzkästlein mit Perlen, aber auch mit Kreuzen und Ketten … In: Bausch, Richard; Christ Herbert u. a. (Hrsg.). *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion: Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Gunter Narr 2003. S. 145-155

Reiss, Mary-Ann. Helping the Unsuccessful Language Learner. In: *The Canadian Modern Language Review,* 39 (2). 1983. S. 257 – 266

Schermelleh-Engel, Karin. *Testtheorien und Testkonstruktion – Gütekriterien*. (Präsentation als PDF-Dokument, Universität Frankfurt am Main) http://user.uni-frankfurt.de/~moosbrug/tttk/Guetekriterien.pdf

Schocker-von Ditfurth, Marita. Die Bedeutung des europäischen Referenzrahmens für die Qualität der Lehr/Lern-Erfahrungen in Schule und Lehrerbildung. In: Bausch, Richard; Christ Herbert u. a. (Hrsg.). *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion: Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Gunter Narr 2003. S. 164-172

Shohamy, Elana. The Power of Tests – A Critical Perspective on the Uses of Language Tests. Pearson Education 2001

Shohamy, Elana; Donitsa-Schmidt, Smadar; Ferman Irit. Test impact revisited: washback effect over time. In: *Language testing,* Vol. 13 (3). 1996. S. 298-317

Tornberg, Ulrika. Das metakognitive Klassenzimmer. Oder: Der lange Weg zurück zum spontanen Lernverhalten. In: *Fremdsprache Deutsch, Sondernummer: Autonomes Lernen*. 1996. S. 24-29

Wall, Dianne. Introducing new tests into traditional systems: insights from general education and from innovation theory. In: *Language testing,* Vol. 13 (3). 1996. S. 334-354

Wall, Dianne. The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? In: *System*, 28. 2000. S. 499-509

Watanabe, Yoshinori. Washback Effects of College Entrance Examinations on Language Learning Strategies. In: *JACET Bulletin,* 23. 1990. S. 175-94

Watanabe, Yoshinori. Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. In: *Language testing,* Vol. 13 (3). 1996. S. 318-333

Watanabe, Yoshinori. Does the University entrance examination motivate learners?: A case study of learner interviews. In: Akita Association of English Studies (Hrsg.). *Trans-equator changes: A collection of academic papers in honour of Professor David Ingram.* 2001. S. 100-110

Watanabe, Yoshinori. (2004a) Methodology in Washback. In: Cheng, Liying; Watanabe, Yoshinori; Curtis, Andy. *Washback in Language Testing – Research Contexts and Methods.* Lawrence Erlbaum Associates 2004. S. 19-36

Watanabe, Yoshinori. (2004b) Teacher Factors Mediation Washback. In: Cheng, Liying; Watanabe, Yoshinori; Curtis, Andy. *Washback in Language Testing – Research Contexts and Methods.* Lawrence Erlbaum Associates 2004. S. 129-146

Wenden, Anita. *Learner Strategies for Learner Autonomy*. New York: Prentice Hall 1991

Werff, Frauke von der; Gerbes Johannes. Externe Zertifizierung für Schüler – Eine Erfolgsstory aus Italien. Wie Prüfungen den Deutschunterricht verändern können. In: *Fremdsprache Deutsch*, Heft 34/2006 („Kompetenzen testen, prüfen, zertifizieren"). Klett 2006. S. 40-44

Japanischsprachige Quellen

Fujiwara, Mieko. „Gengo kyôiku Yôroppa no furêmuwâku: Europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen" no hyôka wo megutte. In: Language and culture: the journal of the Institute for Language and Culture, Vol.8. Kônan University. 2004. S. 107-124
(藤原三枝子：『言語教育ヨーロッパのフレームワーク：Europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen』の評価をめぐって)

Japan Foundation. 『ヨーロッパにおける日本語教育事情とCommon European Framework of Reference for Languages』(kann unter http://www.jpf.go.jp/j/japan_j/publish/euro/ vollständig runtergeladen werden).

NakajimaYoshimichi. Taiwa no nai shakai, PHP shinsho 1997
(中島義道：対話のない社会)

Sekiguchi Ichirô. 'Manabu' kara 'tsukau' gaikokugo e, Shûeisha shinsho 2000
(関口一郎：「学ぶ」から「使う」外国語へ―慶応義塾藤沢キャンパスの実践)