Published in Bulletin VALS-ASLA (Vereinigung für angewandte Linguistik in der Schweiz) 84, 191-227, 2006
which should be used for any reference to this work

Überlegungen zur Sprachkompetenzbeschreibung und Testvalidierung im Projekt HarmoS/Fremdsprachen

Peter LENZ
Lern- und Forschungszentrum Fremdsprachen, Universität Freiburg/CH,
Criblet 13, CH-1700 Fribourg
peter.lenz@unifr.ch

The Swiss Conference of Cantonal Ministers of Education (EDK/CDIP) intends to harmonize the cantonal school systems. One of the means to achieve this goal is to define educational standards for a number of school subjects, among them French, German and English as foreign languages (the exact combination of L2's varies between the parts of the country). The EDK/CDIP has entrusted consortiums of experts with this task. The article first describes the exact mission of the L2 Consortium in more detail and provides an overview of the specific conditions in which it is carried through. The article then focuses on issues regarding models of foreign-language proficiency, as well as the operationalisation of such models as tests. Fundamental considerations on practicable models and research design are given special emphasis; at the same time information is provided on the concrete work undertaken by the L2 Consortium.

Keywords:
Educational standards, foreign languages, language competence and proficiency, language testing, *Evidence-centred Design*

1. Ausgangslage

Im Jahr 2001 initiierte die Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) das Projekt "Harmonisierung der obligatorischen Schule" (HarmoS)[1]. Ziel ist es, die 26 kantonalen Schulsysteme der Schweiz, wenn nicht zu vereinheitlichen, dann doch in bestimmten Belangen zu harmonisieren. Die Arbeiten laufen gleichzeitig auf einer politischen und auf einer wissenschaftlichen Ebene: Einerseits sollen durch einen neuen verbindlichen Staatsvertrag die Kantone zur Harmonisierung der obligatorischen Schule verpflichtet werden, andererseits sollen wissenschaftliche Konsortien Vorschläge für nationale Bildungsstandards ausarbeiten, die eine wichtige Grundlage für die Steuerung des harmonisierten Bildungssystems bilden sollen (als Referenzpunkt nicht nur für ein nationales Bildungsmonitoring, sondern auch für Planungsarbeiten, zum Beispiel Lehrplanarbeit, auf regionaler und lokaler Ebene). Ziel der EDK ist die

[1] Mehr Informationen zu HarmoS finden sich unter der Adresse www.edk.ch > Tätigkeitsbereiche > HarmoS.

Unterzeichnung des Staatsvertrags im Jahre 2007; Anfang 2008 sollen die Expertenvorschläge zu den Bildungsstandards vorliegen, damit sie anschliessend den politischen Prozess bis zur Verabschiedung durchlaufen können.

Vorschläge für Bildungsstandards werden gegenwärtig für vier Fächer bzw. Fächerbereiche ausgearbeitet: L1 (lokale Landessprachen Deutsch, Französisch, Italienisch); Fremdsprachen (Landessprachen Französisch und Deutsch, dazu Englisch); Mathematik und Naturwissenschaften. Die EDK erwartet von den vier Fach-Konsortien in einer ersten Phase (Arbeiten zwischen 2005-2007) einen theoretisch, didaktisch und so weit wie möglich empirisch abgestützten Vorschlag für ein Kompetenzmodell mit Kompetenzstufen, sowie einen Expertenvorschlag für Basisstandards für die 6. und 9. und teilweise auch für die 2. Klasse (betrifft die Fremdsprachen nicht). Vom Fremdsprachenkonsortium wird für die Schulfremdsprachen Französisch, Deutsch und Englisch ein grundsätzlich einheitliches Kompetenzmodell mit der Möglichkeit von unterschiedlichen Akzentuierungen bei der Umsetzung erwartet. Ausgangspunkt ist das vorgegebene Fächersystem; es wird also nicht erwartet, dass ein Akzent auf eine einzelfachübergreifende Kompetenz in Fremdsprachen gelegt würde. Andere Fremdsprachenfächer als die drei genannten könnten zwar auf Initiative von speziell interessierten Kantonen oder Regionen (z.B. Graubünden) an das Hauptprojekt angeschlossen werden, doch gehören diese Fächer nicht zum Auftragsumfang.

Das Modell und der Vorschlag für Standards sollen einerseits beschrieben und andererseits in Form von empirisch erprobten Aufgaben operationalisiert werden. Für die empirische Validierung von Modell und Aufgaben im Frühjahr 2007 (Hauptuntersuchung) stehen sprachregional repräsentative Schülerstichproben für einen schriftlichen Test (*paper and pencil*) zur Verfügung. Die Konsortien haben zudem die Möglichkeit, mit finanzieller Unterstützung der EDK weiterführende Untersuchungen, so genannte "Feldtest", durchzuführen. Für die Planung und Auswertung der empirischen Untersuchungen steht eine Methodologiegruppe als Ansprechpartner zur Verfügung.

Im Folgenden wird das HarmoS-Fremdsprachenkonsortium kurz vorgestellt (Kap. 2) und sein Auftrag präzisiert (Kap. 3). Den eigentlichen Hauptteil bilden grundsätzliche Überlegungen zu Modellen von fremdsprachlicher Kommunikationsfähigkeit (Kap. 4), zur Operationalisierung von solchen Modellen für Tests und Untersuchungen (Kap. 5) sowie zur Testentwicklung und -validierung (Kap. 6). Zum Schluss (Kap. 7) wird das Untersuchungsdesign in HarmoS, in Anlehnung an die vorhergehenden Kapitel kurz skizziert, soweit dies zum jetzigen Zeitpunkt möglich ist.

2. Das Konsortium Fremdsprachen

Vertragsnehmer und direkter Ansprechpartner der EDK ("Leading house") ist das Lern- und Forschungszentrum Fremdsprachen der Universität Freiburg. Partner sind – in der Reihenfolge ihres Engagements – die Pädagogische Hochschule Zürich, die Universität Bern, die Pädagogische Hochschule Freiburg, das Institut de recherche et de documentation pédagogique (IRDP) in Neuchâtel, die Pädagogische Hochschule Zentralschweiz in Luzern sowie die Pädagogische Fachhochschule Graubünden in Chur. Bemerkenswert ist, dass bei allen beteiligten Institutionen die Eigenmittel, die sie in das Projekt einbringen, (zum Teil beträchtlich) grösser sind als die finanzielle Unterstützung durch die EDK.

Das Konsortium wird von einem Fach-Beirat begleitet. Für die Durchführung der Tests zum Sprechen (mündliche Produktion und Interaktion), sowie die Korrektur und Bewertung der Schülerarbeiten werden temporäre Hilfskräfte beigezogen; für die Vorerprobungen der Aufgaben sowie die erwähnten Feldtests bestehen auch Kontakte zu Lehrpersonen und deren Klassen in verschiedenen Regionen.

3. Zum Auftrag des Konsortiums

3.1 Ein Kompetenzmodell mit Kompetenzstufen

Wie oben erwähnt besteht der Auftrag des wissenschaftlichen Konsortiums darin, einen theoretisch, didaktisch und so weit wie möglich empirisch abgestützten Vorschlag für ein Kompetenzmodell mit Kompetenzstufen zu machen, sowie darauf basierend einen Expertenvorschlag für Basisstandards für die Nahtstellen zwischen der Primarstufe und der Sekundarstufe I sowie zwischen der Sekundarstufe I und der Sekundarstufe II zu formulieren.

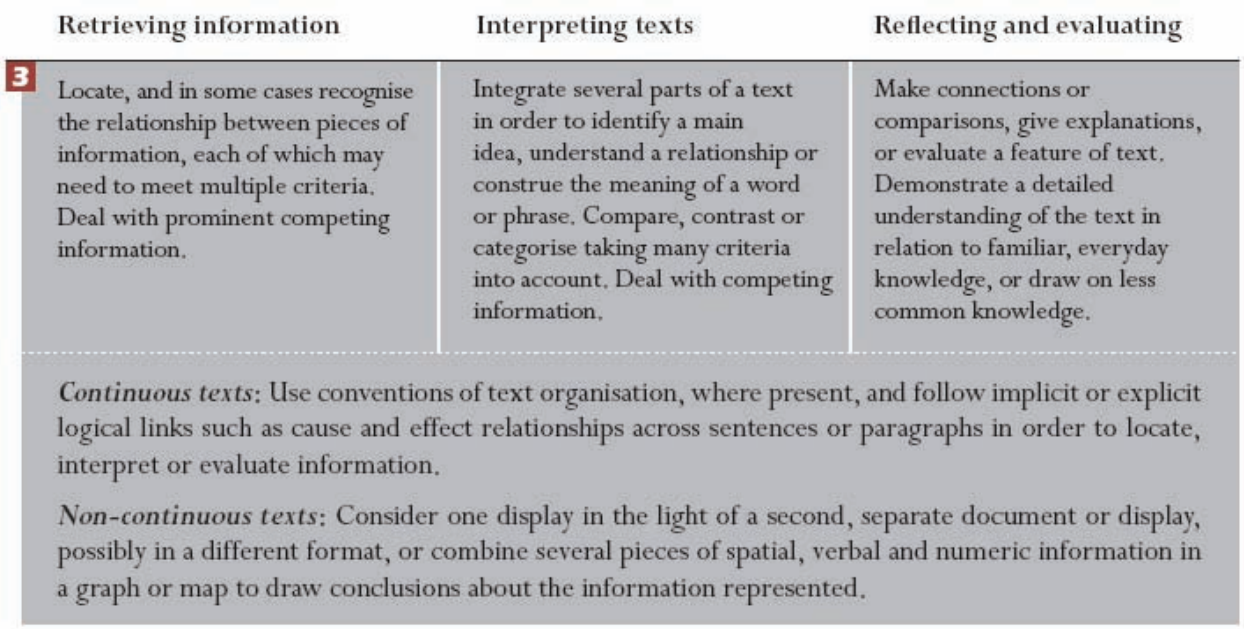
Als Beispiel für ein Kompetenzmodell mit Kompetenzstufen, das sich auf die Ergebnisse empirischer Untersuchungen stützt, kann das Modell genommen werden, das aus den PISA-Untersuchungen der OECD zur *reading literacy* konstruiert wurde: In Bezug auf das Kompetenzmodell ergab sich aus den Ergebnissen, dass drei Leseprozesse unterschieden werden können (*retrieving information; interpreting; reflecting and evaluating*); zudem waren die Ergebnisse in Bezug auf kontinuierliche und diskontinuierliche Texte (z.B. Tabellen) unterschiedlich. Die Items (d.h. die Einzelaufgaben, die jeweils ein Kreuz, ein Wort oder einen mehr oder weniger langen Text als Antwort verlangen) in der Untersuchung wurden alle hinsichtlich der relevanten Merkmale in diesen Kategorien kodiert, sodass eine *item map* erstellt werden konnte, in der auf einen Blick ersichtlich wird, welche Items welchem Leseprozess und welchem Textformat zugerechnet werden:

○ Types of Process (Aspect) ■ Text Format Composite item map	Types of Process (Aspect)			Text Format	
	Retrieving Information	Interpreting	Reflecting and evaluating	Continuous	Non- continuous
822: HYPOTHESISE about an unexpected phenomenon by taking account of outside knowledge along with all relevant information in a COMPLEX TABLE on a relatively unfamiliar topic. (score 2)			○		■
727: ANALYSE several described cases and MATCH to categories given in a TREE DIAGRAM, where some of the relevant information is in footnotes. (score 2)		○			■

Abb. 1 Auszug aus einer *item map* von PISA (OECD 2003, 123)

Dazu ein Beispiel als Lesehilfe: Bei einem Item mit der (relativ hohen) Schwierigkeit 822 (auf einer Skala mit Mittelpunkt 500) müssen Hypothesen aufgestellt und in einer komplexen Tabelle überprüft werden (usw.); Hypothesen aufstellen und überprüfen wird der Prozessdimension Reflektieren und Evaluieren zugerechnet, die komplexe Tabelle den nicht-kontinuierlichen Texten.

Die fünf Kompetenz- bzw. Leistungsniveaus, die in der internationalen Schülerpopulation unterschieden werden konnten, wurden unter Rückgriff auf die solchermassen kodierten Items beschrieben. Dank der relativ grossen Itemzahl war es möglich mit einiger Sicherheit Aussagen zu machen, die von den einzelnen Items abstrahieren:



	Retrieving information	Interpreting texts	Reflecting and evaluating
3	Locate, and in some cases recognise the relationship between pieces of information, each of which may need to meet multiple criteria. Deal with prominent competing information.	Integrate several parts of a text in order to identify a main idea, understand a relationship or construe the meaning of a word or phrase. Compare, contrast or categorise taking many criteria into account. Deal with competing information.	Make connections or comparisons, give explanations, or evaluate a feature of text. Demonstrate a detailed understanding of the text in relation to familiar, everyday knowledge, or draw on less common knowledge.

Continuous texts: Use conventions of text organisation, where present, and follow implicit or explicit logical links such as cause and effect relationships across sentences or paragraphs in order to locate, interpret or evaluate information.

Non-continuous texts: Consider one display in the light of a second, separate document or display, possibly in a different format, or combine several pieces of spatial, verbal and numeric information in a graph or map to draw conclusions about the information represented.

Abb. 2 Beschreibung des mittleren von fünf Niveaus der PISA Reading-Literacy-Studie (OECD 2003, 127)

Im Bereich der Fremdsprachen liegt bereits das Kompetenzmodell des *Gemeinsamen europäischen Referenzrahmens* (Europarat, 2001) vor, das allerdings nicht aufgrund von Tests, sondern durch die Anwendung von Kompetenzbeschreibungen auf Lernende entwickelt wurde (vgl. z.B. North & Schneider, 1998; Schneider & North, 2000; North, 2000). Der *Referenzrahmen* unterscheidet zwischen sechs "Referenzniveaus" (*A1* bis *C2*), die ein sehr breites Kompetenzspektrum abdecken ("elementare Sprachverwendung" bis "kompetente Sprachverwendung"). Diese Niveaus werden durch eine Reihe von skalierten Kompetenzbeschreibungen illustriert. Für unterschiedliche Verwendungszwecke wurden solche Kompetenzbeschreibungen (Deskriptoren) in verschiedenen Instrumenten zusammengefasst, z.B. in der "Globalskala" (Europarat, 2001: 35), die einem breiten Publikum eine rasche Orientierung im Niveausystem vermitteln soll, im "Raster zur Selbstbeurteilung" (Europarat, 2001: 36), der den Lernenden selbst eine Grobeinstufung im System hinsichtlich der fünf Fertigkeitsbereiche Hören, Lesen, interaktives und zusammenhängendes Sprechen sowie Schreiben erlaubt, oder auch im "Beurteilungsraster zur mündlichen Kommunikation" (Europarat, 2001: 37f.), der verwendet werden kann um mündliche Leistungen hinsichtlich von relevanten Kategorien (wie dem Spektrum der sprachlichen Mittel, Korrektheit oder Flüssigkeit) zu beurteilen und in Bezug auf die Referenzniveaus einzustufen.

Das Modell des *Referenzrahmens* wurde im Rahmen des Projekts *IEF*[2] der Deutschschweizer Kantone (vgl. Lenz & Studer, 2004; Materialien in: NW EDK, 2007) spezifischer an die Bedürfnisse und an das Niveauspektrum der Schülerinnen und Schüler von ca. 11 bis 16 Jahren angepasst. Weiter wurden auch Testaufgaben entwickelt und erprobt, die sich an die Referenzniveaus und deren Beschreibungen anlehnen. Aus verschiedenen Gründen ist es sinnvoll – und wird von der Erziehungsdirektorenkonferenz als Auftraggeber auch erwartet –, dass sich das in HarmoS zu schaffende Stufenmodell explizit auf den *Referenzrahmen* und die im Rahmen von *IEF* geschaffenen Ergänzungen bezieht. Insbesondere wäre es wenig sinnvoll, mit den Mitteln des Bildungsmonitorings und der Bildungsstandards (und den künftig darauf bezogenen Lehrplänen) ohne zwingende Gründe und exklusiv für die Schweiz ein konkurrierendes Referenzsystem aufzubauen, während sich das Referenzsystem des Europarates auf europäischer wie auch auf schweizerischer Ebene gerade erst als "gemeinsame Währung" für die Verständigung über Sprachkompetenzen durchsetzt. Viel eher ist es angezeigt, durch die Nutzung aller möglichen Synergien auf das gemeinsame Ziel einer Verbesserung des Sprachunterrichts und letztlich der fremdsprachlichen Kommunikationsfähigkeit der Schüler/innen hinzuarbeiten.

3.2 Basis-outcome-Standards

Die EDK hat sich für eine Standardisierung mittels Basisstandards entschieden. Anders als bei Regelstandards (durchschnittlich erreichte Kompetenz in einer Population) oder Maximalstandards (Ideal) wird bei Basisstandards klar beschrieben, was als Minimum von möglichst allen erreicht werden soll (vgl. Klieme *et al.*, 2003: 27). Unter der zu erreichenden Basis wird nicht nur ein bestimmtes Niveau verstanden, sondern auch ein bestimmter inhaltlicher Kernbereich eines Fachs.

Das zweite wichtige Merkmal der HarmoS-Standards besteht darin, dass *outcome-* oder *performance standards* beschrieben werden und überprüfbar gemacht werden sollen; sie sollen sich also auf Lern*ergebnisse* beziehen, und nicht wie *content standards* auf zu behandelnde Inhalte und Lehrpläne oder wie *opportunity-to-learn standards* auf Lerngelegenheiten und -erfahrungen.

Bildungsstandards – insbesondere als *outcome* oder *performance standards* – dienen in erster Linie der Steuerung des Bildungssystems. Es gehört mit zu den zentralen Aufgaben von Testverfahren, die auf Basisstandards bezogen

---

[2] IEF: "Instrumente für die Evaluation von Fremdsprachenkompetenzen". Das Projekt IEF wurde von der Bildungsplanung Zentralschweiz initiiert und geleitet (Monika Mettler) und von den Deutschschweizer Kantonen finanziert. Projektbearbeitung (2002-2006): Lern- und Forschungszentrum Fremdsprachen der Universität Freiburg/CH (Peter Lenz, Thomas Studer und Eva Wiedenkeller). Der grössere Teil der in IEF entwickelten Materialien wird unter der Bezeichnung *lingualevel* (NW EDK, 2007) vertrieben.

sind, sichtbar zu machen, an welchen Stellen welche Art von Unterstützung für die Leistungserbringer (lokale Bildungsplanung und -administration, Schulen) nötig ist, damit die in den Standards beschriebenen Leistungen von den Schülerinnen und Schülern im erwarteten Ausmass erreicht werden. Einheitliche *outcome standards* rühren nicht daran, dass es weit gehend den Leistungserbringern überlassen ist, auf welche Weise sie die Schülerinnen und Schüler zu den Standards hinführen, und auch wie sie die nicht-standardisierten Bereiche gestalten.

3.3 Entwicklungsszenarien für den Fremdsprachenunterricht

Für die Fremdsprachenfächer ergibt sich eine gewisse Komplikation aus dem Umstand, dass die Szenarien für den Fremdsprachenunterricht in der Schweiz im Wandel begriffen sind. Die Beschlüsse der Erziehungsdirektoren vom 25. März 2004 (EDK, 2004) sehen im Wesentlichen eine Verstärkung und Harmonisierung in zwei Schritten bzw. "Szenarien" vor:

- Szenario 2006: Alle Schülerinnen und Schüler lernen spätestens ab dem 5. Schuljahr mindestens eine zweite Landessprache. Der Englisch-unterricht ab dem 7. Schuljahr wird für alle Schülerinnen und Schüler eingeführt.
- Szenario 2010/12: Spätestens bis zum 5. Schuljahr setzt der Unterricht in mindestens zwei Fremdsprachen (davon mindestens einer Landes-sprache) ein. Der Unterricht der ersten Fremdsprache beginnt spätestens ab dem dritten Schuljahr, der Unterricht der zweiten Fremdsprache spätestens ab dem fünften Schuljahr.

Die empirische Hauptuntersuchung und die Vorschläge für Basisstandards in den Fremdsprachenfächern müssen also zu einem Zeitpunkt erfolgen, zu dem die im März 2004 beschlossen Szenarien für die Entwicklung des Fremdsprachenunterrichts bei Sechst- und Neuntklässlern noch kaum umgesetzt sind. Von den im Jahr 2007 empirisch ermittelbaren Kompetenzen sind folglich nur vorsichtige Prognosen möglich im Hinblick auf Standards für Schülerinnen und Schüler, die unter anderen Bedingungen (oft früherer Beginn, teilweise neue Unterrichtsformen und Lernmaterialien) Französisch, Deutsch oder Englisch gelernt haben werden. Immerhin ist aber zu erwarten, dass hinsichtlich der Überprüfung von Aspekten des Kompetenzmodells und der Aufgabenvalidierung keine wesentlichen Abstriche gemacht werden müssen. Denn Erkenntnisse in diesen Bereichen wären nur dann nicht in die Zukunft übertragbar, wenn die zukünftigen Schülerinnen und Schüler nicht nur besser oder weniger gut wären oder andere Kompetenzprofile hätten, sondern wenn sie ganz anders funktionieren würden.

3.4 Beteiligung des Fremdsprachenkonsortiums an der HarmoS-Hauptuntersuchung

An sich würde es sich aus der Perspektive der Systemsteuerung, aber auch der Fremdsprachendidaktik und Sprachlehrforschung, geradezu aufdrängen zum jetzigen Zeitpunkt eine umfassende Momentaufnahme der vorhandenen Kompetenzen zu machen, damit bei weiteren Untersuchungen zu einem späteren Zeitpunkt, wenn die neuen Szenarien in der Praxis einmal gegriffen haben, ein solider Bezugspunkt vorliegen würde, um die Effekte der neuen Szenarien zu erfassen. Ein Jahr nach Projektbeginn, also im Herbst 2006, machten die Erziehungsdirektoren aber klar, dass sie den in Abschnitt 3.3 erwähnten Einwand, dass von den jetzt erfassbaren Schülerkompetenzen nur indirekt Standards für die Zukunft abgeleitet werden könnten, für schwerwiegend genug hielten, um den Umfang der Stichprobe, die dem Fremdsprachenkonsortium zur Verfügung stehen sollte, drastisch einzuschränken. Als Ersatz wurde eine grosse, repräsentative Untersuchung für die Jahre nach dem Greifen des Szenarios 2010/12 in Aussicht gestellt.

Nach Verhandlungen mit der EDK wurde dem Fremdsprachenkonsortium aber in den Untersuchungen von März/April 2007 doch Platz eingeräumt: in den 9. Klassen der Deutschschweiz und der Westschweiz je 560 Minuten Testzeit (insgesamt zur Verfügung stehende Zeit für die einmalige Bearbeitung aller verschiedenen Aufgaben)[3]; darin müssen die Aufgaben zu zwei Fremdsprachen untergebracht werden; in den Untersuchungen in den 6. Klassen der zwei Landesteile steht jeweils die Hälfte der Zeit (280 Minuten) zur Verfügung; allerdings muss darin auch nur eine Sprache (nämlich die zweite Landessprache) Platz finden. Die zur Verfügung stehende Testzeit wird voraussichtlich genügen, um Schülerkompetenzen über ein angemessen breites Niveauspektrum hinweg zu erfassenden und auch um genügend Aufgaben validieren zu können, die in späteren Untersuchungen als Anker zur jetzigen Untersuchung dienen können, sodass quantitative Vergleiche zwischen den Schülerkompetenzen zu verschiedenen Testzeitpunkten möglich sein werden. Dagegen wird es schwierig sein, differenzierte Einsichten hinsichtlich eines Kompetenzmodells – jenseits der primären Unterscheidungen vor allem zwischen Lese- und Schreibkompetenz – empirisch genügend zu untermauern, weil einzelne Phänomene nicht in genügend verschiedenen Aufgaben zu beobachten sein werden, um Verallgemeinerungen zu erlauben. Die 6. und 9. Klassen bekommen zum Teil identische Aufgaben, sodass (voraussichtlich) eine kontinuierliche Skala

---

[3] Für die 9. Klasse in der Deutschschweiz bedeutet dies zum Beispiel konkret: Französisch: 3-mal 30 min. und 2-mal 20 min. Lesen; 1-mal 30 min. und 5-mal 20 min. Schreiben; 20 min. C-Test. Englisch: 2-mal 30 min., 2-mal 20 min. und 1-mal 10 min. Lesen; 1-mal 30 min. und 6-mal 20 min. Schreiben.

entwickelt werden kann, auf welcher den Kompetenzzuwachs über einige Jahre hinweg dargestellt werden kann. Ebenso werden die Aufgaben für den Gebrauch in der Deutsch- und in der Westschweiz sowie für das Testen der verschiedenen Zielsprachen Französisch, Deutsch und Englisch teilweise übersetzt, sodass für Quervergeiche zumindest eine gute Ausgangsbasis besteht. Aus statistischen Gründen (Messfehler; Vergleichbarkeit der Mittelwerte der Teilstichproben) wird jedes Testheft (d.h. auch jede Aufgabe) von mindestens 150 zufällig ausgewählten Schülerinnen und Schülern bearbeitet. Bei der Rotation der Testhefte wird darauf geachtet, dass ein Individuum jeweils nur in einer Fremdsprache getestet wird; unterschiedliche Kombinationen mit Aufgaben zur lokalen Schulsprache (L1) kommen dagegen vor.

Unter den gegebenen Umständen, vor allem wegen des enormen Zeitdrucks, aber auch wegen der eingeschränkten Gesamttestzeit in der repräsentativen Hauptuntersuchung, entschied sich das Konsortium für eine Beteiligung nur in den Kompetenzbereichen des Leseverstehens und des Schreibens, nicht aber des Hörverstehens, was gerade hinsichtlich der Standards für die 6. Klasse bedauerlich ist. Zusätzlich wird pro Zielsprache ein einzelner C-Test in die Untersuchung eingebracht. Der C-Test ist eine spezielle Form von Lückentest, bei der jeweils die zweite Worthälfte jedes zweiten Wortes gelöscht wird. Ihm wird, je nach Autor und Population, nachgesagt, dass er sehr effizient und trotzdem recht zuverlässig die generelle Kompetenz in einer Sprache bzw. vor allem die produktive schriftliche Kompetenz voraussagen kann (mehr Informationen zum C-Test unter www.c-test.de). In dieser ersten Untersuchung soll vor allem das Funktionieren des C-Tests in diesen konkreten Populationen, sowie die Korrelation zwischen Schreibkompetenz und C-Test untersucht werden. Zum Hörverstehen sind ausserhalb der offiziellen Hauptuntersuchung kleinere Untersuchungen vorgesehen, die, gerade was Erkenntnisse hinsichtlich eines Kompetenzmodells angeht, eher mehr Möglichkeiten bieten als die Hauptuntersuchung, bei der keine direkten Klassenkontakte – und damit zum Beispiel keine vertiefenden qualitativen Untersuchungen – möglich sind. Einzig repräsentative Erkenntnisse über die Verteilung der Hörverstehenskompetenzen in der Bevölkerung werden nicht möglich sein. Untersuchungen zum Sprechen waren wegen des aufwendigeren Verfahrens im Rahmen der Hauptuntersuchung nie vorgesehen. Geplant sind aber Tests und Videoaufnahmen mit insgesamt 240 Schüler/innen in der Deutsch- und in der Westschweiz (für Französisch bzw. Deutsch in der 6. Klasse; für Französisch und Englisch bzw. Deutsch und Englisch in der 9. Klasse). Mithilfe dieser Aufnahmen soll für verschiedene Sprachen und Kontexte das Spektrum der vorhandenen Kompetenzen grob illustriert werden. Zudem können Testaufgaben und Testmodelle erprobt sowie konkret darauf bezogene Beurteilungskriterien validiert werden.

4. Der Gegenstand der Beschreibung und Standardisierung: "Sprachkompetenz"

4.1 Problemlösungsfähigkeit

Wie oben erwähnt sieht der Auftrag der EDK eine Verbindung von Kompetenzmodell und Standards vor. Es soll ein Kompetenzmodell entwickelt und (teilweise) empirisch validiert werden, auf dessen Grundlage die Anforderungen an die Lernenden verschiedener Stufen bzw. die Kompetenzen dieser Lernenden kommuniziert werden können. Die grundlegenden Erwartungen der EDK an ein Kompetenzmodell basieren auf der Expertise, die als Grundlagenpapier für Bildungsstandards der deutschen Kultusministerkonferenz (KMK) verfasst wurde (Klieme *et al.*, 2003). In dieser Studie wird "Kompetenz" – in Anlehnung an Weinert – folgendermassen definiert:

> […] die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können. (Klieme *et al.*, 2003: 21)

Kompetenz wird damit im Wesentlichen als (kognitive) Problemlösungsfähigkeit verstanden, die in Abhängigkeit von bestimmten "Bereitschaften und Fähigkeiten" eingesetzt wird. Mit Klieme als Referenz ist einigermassen klar vorgegeben, dass bei den zu schaffenden Standards und den zu messenden *outcomes* die Fähigkeit, mithilfe von Sprache Probleme zu lösen bzw. Kommunikationsaufgaben zu erfüllen, eine zentrale Stellung einnehmen wird. Damit ist nicht ausgeschlossen, dass auch die genannten "Bereitschaften und Fähigkeiten" selbst zum Gegenstand gemacht werden (z.B. Messung der Motivation für Französisch). Das Fremdsprachenkonsortium hat sich aber entschieden, zumindest bei der Operationalisierung des Modells in der Hauptuntersuchung auf die Problemlösungsfähigkeit zu fokussieren – konkret auf die Fähigkeit, sprachlich-kommunikative Aufgaben zu lösen.

4.2 Zur Modellierung von fremdsprachlicher Kommunikationsfähigkeit

4.2.1 Von Hymes zu Bachman

Der Paradigmenwechsel in der Linguistik und später in der Fremdsprachendidaktik vor allem der siebziger Jahre – weg von strukturalistisch oder behavioristisch inspirierten Ansätzen, hin zu einem kommunikativen, auf der Pragmalinguistik aufbauenden Kompetenzverständnis – verlangte nach umfassenderen Modellen fremdsprachlicher Kompetenz, als dies vorher der Fall war, weil der Aspekt der Sprachverwendung dazukam und in den Vordergrund rückte. Im Kontext des *language testing* ist die theoretische Auseinandersetzung um ein umfassendes und gleichzeitig noch praktisch umsetzbares Modell fremdsprachlicher Kommunikationsfähigkeit gut

dokumentiert. Besonders hervorzuheben sind die Darstellungen von McNamara (1996: 48-90) und North (2000: 41-129). Der eigentliche Knackpunkt der ganzen Diskussion war, was Hymes (1972: 282f.) in seinem Modell mit *ability for use* (Sprachverwendungskompetenz – neben *(tacit) knowledge* und *performance*) bezeichnete. Wie könnte es möglich sein, in einem Sprachkompetenzmodell nicht spezifisch sprachliche Aspekte wie (metakommunikative) Strategien der Handlungssteuerung, Affekte oder die sozialen Merkmale von Kommunikation zu berücksichtigen, ohne durch zu grosse, nicht kontrollierbare Komplexität die praktische Umsetzbarkeit besonders im Bereich des Sprachentestens zu gefährden? In einer sehr anschaulichen Metapher vergleicht McNamara (1996) die Zurückhaltung einiger Autoren gegenüber der *ability for use* mit der Furcht vor dem Öffnen der mythischen Büchse der Pandora. Canale and Swain (1980) sind ein Beispiel dafür; sie verzichten darauf, *ability for use* als Komponente oder Aspekt in ihr Modell aufzunehmen und bleiben bei einem Kompetenz-Performanz-Dualismus: *knowledge* (grammatische, soziolinguistische und strategische Wissens- und Könnensressourcen, die zusammen *communicative competence* ausmachen) versus *communicative performance*. Sie fokussieren dabei ihr Interesse einseitig auf die Kompetenz, d.h. auf die (meist unbewussten) sprachlichen Fähigkeiten oder Dispositionen, und integrieren *ability for use* als Teil der Performanz, des "blossen" Handelns, nicht in ihr Modell. Damit verzichten sie darauf, die Interaktion der Kompetenzkomponenten (oder -aspekte[4]) in der Performanz zu beschreiben oder dies gar, als das Wesentliche, in den Mittelpunkt zu stellen (vgl. McNamara, 1996: 62f.).

Einen wichtigen Beitrag zur Modellentwicklung insbesondere im Hinblick auf das Testen – durchaus in der angewandt-linguistischen Tradition von Hymes – leisteten in den 1990-er Jahren Bachman und Palmer (Bachman, 1990; Bachman & Palmer, 1996) mit ihrem *interactional model of language test performance.* Das erweiterte Modell von 1996 bezieht nicht nur *ability for use* mit ein, sondern gibt auch affektiven Faktoren eine wichtige Funktion:

[4] North (2000: 68) weist darauf hin, dass es zu einem Zeitpunkt, wo noch wenig gesichertes Wissen über die tatsächlichen Kompetenzkomponenten vorhanden sei, weniger verfänglich wäre, von "Aspekten" anstatt von "Komponenten" zu sprechen, weil dadurch keine Existenzaussagen gemacht würden.

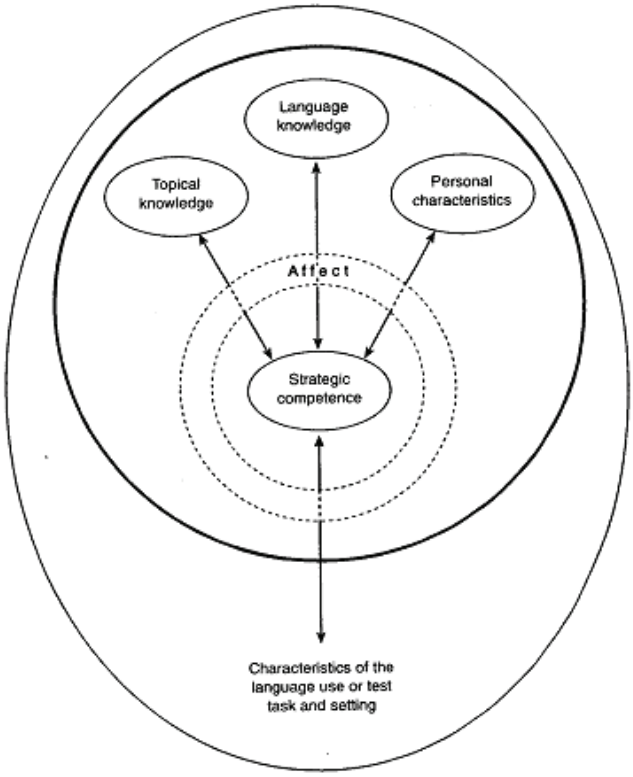


Figure 4.1: Some components of language use and language test performance

Abb. 3 Die wichtigsten Komponenten eines Modells der Sprachverwendung von Bachman & Palmer (1996: 63)

Die *ability for use* wird in Form der "strategischen Kompetenz" in dieses Modell eingeführt. Strategische Kompetenz wird verstanden als:

> "a set of metacognitive components or strategies, which can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities" (Bachman & Palmer, 1996: 70).

Diese metakognitiven Komponenten wirken in drei Hauptbereichen der Handlungssteuerung: Zielbestimmung, Beurteilung und Planung. Ob und in welcher Art gehandelt wird, hängt aber nicht nur vom Vorliegen der notwendigen Kompetenzen ab, sondern auch von Affekten, die gemäss diesem Modell – in Abhängigkeit von Vorerfahrungen – wesentlich mitbestimmen, ob überhaupt und mit wie viel Flexibilität auf eine Kommunikationssituation eingetreten wird. Die Strategien interagieren mit allen übrigen Komponenten des Modells. Diese bestehen zum einen aus Merkmalen des handelnden Subjekts (grosser Kreis mit dicker, durchgezogener Linie im Schema) und zum anderen aus den Merkmalen der Aufgabe und der Situation (unten in der äussersten Ellipse), in der gehandelt wird – Testsituation eingeschlossen (Merkmale von Umgebung,

Aufgabenstellung, Input, erwarteter Antwort; Beziehung zwischen Input und erwarteter Antwort)[5]. Die Merkmale des handelnden Subjekts sind weiter aufgeteilt in drei grosse Bereiche: Weltwissen (*topical knowledge*), Sprachkompetenz (*language knowledge*) und Persönlichkeitsmerkmale (*personal characteristics*).

Das Zusammenwirken von *language knowledge* und strategischer Kompetenz bezeichnet Bachman als *language ability*. Für Bachman & Palmer (1996: 67) ist *Language ability* das Kernelement einer sprachlich-kommunikativen Handlungsfähigkeit:

> It is this combination of language knowledge and metacognitive strategies that provides language users with the ability, or capacity, to create and interpret discourse, either in responding to task on language tests or in non-test language use.

Als Klassifikation für *language knowledge* schlagen Bachman & Palmer Folgendes vor:

Organisationelle Kompetenz	**Pragmatische Kompetenz**
Grammatische Kompetenz: Wortschatz Syntax Phonologie/Graphologie	Funktionale Kompetenz: ideationelle Funktion manipulative Funktion heuristische Funktionen imaginative Funktionen
Textkompetenz: Kohäsion rhetorische oder konversationelle Komp.	Soziolinguistische Kompetenz: Dialekte/Varianten Register Natürlichkeit oder Idiomatizität der Ausdrücke kulturelle Referenzen und Redeweisen

Abb. 4 Sprachkompetenzen (im engeren Sinn) bei Bachman & Palmer (vgl. Bachman & Palmer, 1996: 68)

Diese Aufgliederung von *language knowledge* durch Bachman und Palmer hat recht breite Akzeptanz gefunden, zumindest als Arbeitsmodell[6], denn für die sachliche Richtigkeit der Untergliederungen (d.h. für Entsprechungen im kognitiven Apparat der Sprachverwendenden) liegt kaum empirische Evidenz vor. Bachman und Palmer selbst haben schon in den achtziger Jahren durch Multi-trait-multi-method-Studien immerhin zeigen können, dass eine Grob-

---

5 Die verschiedenen Komponenten des Modells werden in den Kapiteln 3 (Aufgabe/Situation) und 4 (vor allem *language ability* unter den Merkmalen des handelnden Subjekts) von Bachman & Palmer, 1996 ausführlich dargestellt.

6 Die Darstellung der Sprachkompetenzen im "Gemeinsamen europäischen Referenzrahmen für Sprachen (Europarat, 2001) nimmt zwar eine etwas andere Untergliederung vor als Bachman & Palmer (1996), unterscheidet sich aber in den Komponenten (oder Aspekten) nicht wesentlich (vgl. Kap 5.2 des *Referenzrahmens*). Grundunterteilung im *Referenzrahmen*: Linguistische – soziolinguistische – pragmatische Kompetenz.

unterteilung in organisationelle und pragmatische Kompetenz inhaltlich gerechtfertigt ist (vgl. North, 2000: 65).

Bachman & Palmers *language ability* ist als Bezeichnung für die Fähigkeit erfolgreich zu kommunizieren zumindest im Feld des *language testing* weniger gebräuchlich als *(communicative) language proficiency*. Allerdings ist das Spektrum der Bedeutungen von *language proficiency* recht weit: Es erstreckt sich von reiner (kognitiver) Handlungs*disposition* bis zu dem, was an Konsistenz im kommunikativen Handeln beobachtet werden kann – je nachdem, ob *language proficiency* als Begriff in eine *Trait*[7]-Theorie (Interesse an kognitiven Dispositionen) oder in einen interaktionistischen oder behavioristischen Ansatz (Interesse an Handlungskontexten) eingebettet ist (mehr zu diesen Ansätzen weiter unten). Die Verwendung von *proficiency* im Zusammenhang mit den bekannten *Oral Proficiency Interviews* (OPI) und den ACTFL-Skalen ist zum Beispiel am behavioristischen Ende des Spektrums angesiedelt und bezieht sich auf die beobachtbaren Merkmale von sprachlichen Leistungen[8].

Alderson und Banerjee würdigen in ihrem State-of-the-art-Artikel (vgl. Seite 21) von 2002 das *Bachman model* als

> an influential point of reference, being increasingly incorporated into views of the constructs of reading, listening, vocabulary and so on. […] It remains very useful as the basis for test construction, and for its account of test method facets and task characteristics (Alderson & Banerjee, 2002: 80).

Diese Beurteilung muss vor dem Hintergrund gesehen werden, dass "despite considerable consensus, no universal, validated, theoretical model of either communicative competence or of communicative activities exist or is likely to exist for some considerable time" (North, 2000: 123), dass es aber verschiedene Arbeitsmodelle gibt, unter denen dasjenige von Bachman und Palmer offenbar eines der fruchtbarsten und meistverwendeten ist (vgl. Alderson & Banerjee, 2002: 80).

Andere weisen auf Mängel auch des Bachman-Modells hin, darunter McNamara (1996: 75; 85f.): Er kritisiert, dass es als Basis, insbesondere für das Testen von mündlicher Performanz, nicht genüge, weil es die Komplexität der Interaktionen (von Personen- und Kontextfaktoren), die in der Kommunikation tatsächlich spielten, konzeptuell zu wenig erfasse, weil es im Grunde psychologisch orientiert sei, auf Kompetenzen (*knowledge*) fokussiere und dabei soziale Interaktion nur als statische, soziolinguistische Kompetenzdimensionen einschliesse; auch werde Interaktion einseitig von nur

---

7 *Trait*: am besten zu verstehen als (zumindest momentan stabile) psychologische Eigenschaft wie Intelligenz oder Wortschatzkenntnis.

8 "ACTFL": American Council on the Teaching of Foreign Languages. Die *Proficiency*-Diskussion ist ausführlich dargestellt in McNamara, 1996: 76-79 und North, 2000: 47ff.

einem Beteiligten her (in der Regel dem Testkandidaten) aus einem kognitiven Blickwinkel dargestellt. McNamara schlägt dementsprechend vor, kommunikative Performanz verstärkt unter dem Aspekt der sozialen Interaktion und der Ko-Konstruktion zu betrachten und für das Sprachentesten bestimmte Modelle entsprechend durch Erkenntnisse aus der Ethnografie oder der Konversationsanalyse zu erweitern.

McNamara weist an gleicher Stelle zudem darauf hin, dass es nicht ausreiche, in Modellen potenziell wichtige Kontext- und Personenvariablen (Aufgabe, Kandidat/in, Interlokutor/in; Beurteiler/in, Skala/Kriterien, Performanz; Beurteilung) zu identifizieren, ohne aber deren Einfluss und Bedeutung innerhalb der komplexen Konstellation zu kennen.

Während für McNamara aus der Perspektive des mündlichen *performance testing* die sozialen Interaktionen zuoberst auf der Forschungsagenda stehen, sind es für Alderson & Banerjee (2002: 101) die kommunikativen Testaufgaben (*tasks*) – was bei ihrer Bearbeitung wirklich abläuft; wie bestimmte Aufgabeneigenschaften in bestimmten (Test-)Kontexten mit den unterschiedlichen Eigenschaften der an der Bearbeitung beteiligten Personen (v.a. des Testkandidaten) interagieren[9].

> Understanding the nature of the tasks we present to test takers and how these tasks interact with various features, including the characteristics of different test takers within the testing context, presents the most important challenge for language testers for the next few years. It is a conundrum that we have acknowledged for years but have not yet really come to grips with. (Alderson & Banerjee, 2002: 101)[10]

Innerhalb von HarmoS sollen die von McNamara und Alderson/Banerjee genannten Gesichtspunkten zumindest ansatzweise aufgenommen werden, natürlich ohne dass sie umfassend erforscht werden könnten, denn dafür sind jeweils spezifische Projekte nötig. McNamaras Forderungen betreffen vor allem die Tests zum Sprechen: vor allem eine bewusste und gezielte Variation von Aufgabenstellungen, Personenkonstellationen und Interlokutorenverhalten (mit entsprechendem Training) sowie differenzierte Beurteilungen mithilfe von geeigneten Kriterien. Alderson und Banerjee beziehen sich auf Testaufgaben allgemein. In HarmoS sind ausserhalb der Hauptuntersuchung qualitative Untersuchungen zum Funktionieren von Leseverstehens- und Hörverstehens-aufgaben vorgesehen, zum Beispiel zu den Auswirkungen von einmaligem oder zweimaligem Abspielen von Texten beim Hörverstehen oder zu den

---

9 Wichtige Arbeiten zum Einfluss unterschiedlicher Aufgaben auf die mündliche *Test performance* liegen zum Beispiel von Skehan (1998; 2001) und von Robinson (2001; 2005) vor. Skehan bringt zusätzlich den Faktor unterschiedlicher Kompetenzkonstellationen (*individual differences*) bei verschiedenen Personen bzw. Testkandidat/innen ins Spiel.

10 Für Bachman selbst (2002: 468-71) steht ganz ähnlich "the issue of how assessment tasks affect performance" im Vordergrund. Im Sinne eines "way forward" macht er dazu interessante, ausführliche Vorschläge für die praktische Umsetzung in Forschung und Test design.

tatsächlichen Überlegungen von unterschiedlich guten Lernenden beim Bearbeiten von Leseverstehensaufgaben.

4.2.2 Kompetenzbeschreibung im *Gemeinsamen europäischen Referenzrahmen*

Das Kompetenzmodell des *Gemeinsamen europäischen Referenzrahmens* steht in derselben Tradition wie das Bachman-Modell (vgl. Alderson & Banerjee, 2002: 81). Das Sprachverwendungs- und Sprachlernmodell, der "handlungsorientierte Ansatz" des *Referenzrahmens* bezieht die Strategien der Handlungssteuerung, "die kognitiven und emotionalen Möglichkeiten und die Absichten von Menschen" sowie überhaupt "das ganze Spektrum der Fähigkeiten, über das Menschen verfügen" und beim Handeln einsetzen, grundsätzlich mit ein. Deutlich wird die soziale Einbettung des sprachlichen Handelns hervorgehoben:

> Der 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. (Europarat, 2001: 21)[11]

Ziel des *Referenzrahmens* war es weniger, ein Kompetenzmodell von Grund auf neu zu entwickeln, als vielmehr für die kommunikative Sprachkompetenz in Fremdsprachen Referenzniveaus zu *beschreiben*, was natürlich Kompetenzvorstellungen zumindest impliziert. Das Kernstück bilden die Beschreibungen von "kommunikativen Aktivitäten und Strategien" (Kap. 4.4) und die Beschreibungen von "kommunikativen Sprachkompetenzen" (Kap. 5.2), die zwar beide Sprachverwendung auf unterschiedlichen Niveaus beleuchten, aber jeweils aus einer anderen Perspektive: einerseits aus einer Handlungsperspektive (Was kann man typischerweise auf einem Niveau tun?) andererseits aus einer Qualitätsperspektive (Welche sprachlichen Merkmale weisen kommunikative Leistungen auf den verschiedenen Niveaus normalerweise auf?). Zu beiden Perspektiven wurden Kompetenz-

---

[11] Die Beschreibung von "Sprachverwendung" an zentraler Stelle im "Referenzrahmen" zeigt sehr schön, welche Vielzahl unterschiedlicher Faktoren in die Sprachverwendung hineinspielen – und die Beschreibung von kommunikativer Sprachkompetenz zu einem umfassenden Unternehmen machen: "Sprachverwendung – und dies schließt auch das Lernen einer Sprache mit ein – umfasst die Handlungen von Menschen, die als Individuen und als gesellschaftlich Handelnde eine Vielzahl von *Kompetenzen* entwickeln, und zwar *allgemeine*, besonders aber *kommunikative Sprachkompetenzen*. Sie greifen in verschiedenen *Kontexten* und unter verschiedenen *Bedingungen und Beschränkungen* auf diese Kompetenzen zurück, wenn sie *sprachliche Aktivitäten* ausführen, an denen (wiederum) *Sprachprozesse* beteiligt sind, um *Texte* über bestimmte *Themen* aus verschiedenen *Lebensbereichen* (Domänen) zu produzieren und/oder zu rezipieren. Dabei setzen sie *Strategien* ein, die für die Ausführung dieser *Aufgaben* am geeignetsten erscheinen. Die Erfahrungen, die Teilnehmer in solchen kommunikativen Aktivitäten machen, können zur Verstärkung oder zur Veränderung der Kompetenzen führen." (Europarat, 2001: 21)

beschreibungen entwickelt[12], die einem Teil dieses Beschreibungswerks eine empirische Basis verleihen und das Beschriebene dank der Konkretheit und Vielfalt der Beschreibung besonders für Praktiker besser fassbar machen.

Die Beschreibung der "kommunikativen Sprachkompetenzen" ist Bachmans *language knowledge* ähnlich. Unterschieden werden auf einer ersten Ebene linguistische Kompetenzen, soziolinguistische Kompetenzen und pragmatische Kompetenzen, während Bachman nur in *organizational knowledge* und *pragmatic knowledge* – welches *functional knowledge* und *sociolinguistic knowledge* umfasst – unterteilt. Die wichtigste Abweichung betrifft ein dynamischeres Pragmatikverständnis im *Referenzrahmen*, das auch gemeinhin dem Bereich der Performanz (im Gegensatz zur Kompetenz – vgl. North, 2000: 91) zugewiesene Aspekte umfasst: Bei Bachman erscheint *functional knowledge* nur als (statische) Kenntnis der funktionalen Verwendungs- und Interpretationsmöglichkeiten von sprachlichen Mitteln. Der *Referenzrahmen* ordnet der pragmatischen Kompetenz (Diskurskompetenz, funktionale Kompetenz und Schemakompetenz) zwar auch "Wissensbestände" zu, wie zum Beispiel die Kenntnis von Textkonventionen oder sprachlichen Mitteln im Mikrobereich, bezieht aber in wesentlichen Teilen der Beschreibung die strategische Kompetenz direkt mit ein – am augenfälligsten bei den Makrofunktionen (Argumentieren etc.) und bei den Interaktionsschemata (auch "Szenarien" genannt – z.B. Einkauf von Waren und Dienstleistungen). Die Skalen mit den Beschreibungen von pragmatischen Kompetenzaspekten sind ebenfalls so gefasst, dass sie oft mehr umfassen als klar zuordenbare Wissensbestände. Diskurskompetenz ist durch Skalen zu den folgenden Aspekten illustriert: Flexibilität, Sprecherwechsel, Themenentwicklung, Kohärenz und Kohäsion. Zur funktionalen Kompetenz (Schemakompetenz eingeschlossen) existieren zwei qualitative Skalen: Flüssigkeit[13] und Genauigkeit des Ausdrucks. Zahlreiche weitere Aspekte pragmatischer Kompetenz sind (typischerweise) in den Skalen zu den kommunikativen Sprachaktivitäten (mündliche Interaktion) beschrieben, dort allerdings mit einem deutlicher ganzheitlich-handlungsbezogenen Fokus (z.B. Kategorie "Dienstleistungsgespräche", S. 84: "Kann andere um etwas bitten und anderen etwas geben.").

---

[12] Diese Entwicklungsarbeiten fanden hauptsächlich im Rahmen des Nationalen Forschungsprojekts 33 "Wirksamkeit unserer Bildungssysteme" statt (vgl. North & Schneider, 1998; North, 2000; Schneider & North, 1999; 2000a/b).

[13] Nach North (2000: 89-92) ist Flüssigkeit (*fluency*) eine sehr wichtige Kategorie: Zum einen lassen sich Muttersprachler von Flüssigkeit oft mehr beeindrucken als von allen anderen Aspekten einer Leistung und zum anderen fassen Praktiker/innen oft sehr vieles unter "Flüssigkeit", nämlich a) Leichtigkeit des Zugangs zu den sprachlichen Ressourcen; b) die unter a) genannten Aspekte plus Flexibilität, Genauigkeit, Kohärenz und Art der Themenentwicklung im Diskurs; c) die unter a) und b) genannten Aspekte plus *turntaking*, Kooperieren und Reparieren in der Interaktion (North, 2000: 91).

Wie bereits erwähnt rückt Kapitel 4.4 ("kommunikative Aktivitäten und Strategien") in seinen Beschreibungen den Aspekt des Tun-Könnens in den Vordergrund. Die entsprechenden Handlungsbeschreibungen werden aufgrund wesentlicher Merkmale (Domäne: privat, öffentlich, Ausbildung usw.; Themen; Textsorten) zu Gruppen zusammengefasst. Beispiele solcher Gruppierungen sind beispielsweise im Bereich des Hörens "Gespräche zwischen Muttersprachlern verstehen", Ankündigungen, Durchsagen und Anweisungen verstehen" oder "Radiosendungen und Tonaufnahmen verstehen"[14]. Eine durchgängige Strukturierung des Feldes der kommunikativen Handlungen in dieser Art würde sich nach North (2000: 51f.) vor allem deshalb aufdrängen, weil dank solcher Kategorisierungen die tatsächlich existierenden Kompetenzprofile der Lernenden besser erfasst werden könnten. Im *Referenzrahmen* wird trotzdem auch noch an den klassischen "Fertigkeiten" (Hören, Lesen, Sprechen, Schreiben) festgehalten, indem nämlich die Gruppierungen von kommunikativen Sprachaktivitäten wiederum den Fertigkeiten zugeordnet werden, obschon genau genommen nicht jede relevante Gruppierung genau einer Fertigkeit zugeordnet werden kann. So umfasst zum Beispiel die Kategorie "Texte verarbeiten" (z.B. Texte zusammenfassen) Lesen und Schreiben, ohne dass der Schwerpunkt der Tätigkeit klar einer der beiden Fertigkeiten zuzuordnen wäre.

Im *Referenzrahmen* wird das Fertigkeitenschema im Vergleich mit der klassischen Vierteilung in Hören, Lesen, Sprechen und Schreiben weiter ausgebaut, allerdings können nicht für alle Fertigkeiten auch Kompetenzbeschreibungen zur Verfügung gestellt werden. Die folgende Darstellung vermittelt einen Überblick:

---

[14] Zwei Beispiele solcher Beschreibungen zur Kategorie "Gespräche zwischen Muttersprachlern verstehen: Niveau A2: "Kann im Allgemeinen das Thema von Gesprächen, die in seiner/ihrer Gegenwart geführt werden, erkennen, wenn langsam und deutlich gesprochen wird"; Niveau C1: "Kann komplexer Interaktion Dritter in Gruppendiskussionen oder Debatten leicht folgen, auch wenn abstrakte, komplexe, nicht vertraute Themen behandelt werden".

		Kommunikationsmodus			
		Produktion	**Rezeption**	**Interaktion**	**Sprachmittlung**
Kanal	**mündlich**	✓	✓	✓	___
	schriftlich	**(✓)**[15]	**(✓)**	**(✓)**	

Abb. 5 Das Fertigkeitenschema des *Referenzrahmens* (vgl. Europarat, 2001: 25f.)

Es ist symptomatisch für das Bestreben der Autoren der Kompetenzbeschreibungen und des *Referenzrahmens* etwas praktisch leicht Nutzbares zu schaffen, dass vor allem aus praktischen Überlegungen heraus am Fertigkeitenschema festgehalten wird. Auch bei der Entscheidung für die Art, wie die Sprachaktivitäten mittels Deskriptoren beschrieben werden, spielte die Verstehbarkeit für Praktiker (Unterrichtende, Prüfende) eine zentrale Rolle: Anstelle eines homogenen, aber abstrakten Beschreibungssystems wurden eigenständige Beschreibungen von sprachlichen Handlungen gewählt, die intuitiv gut überschaubar sind und deren Schwierigkeit von kompetenten Sprechern ähnlich eingeschätzt wird, sodass sie geeignet sind, in ihrer Gesamtheit die verschiedenen Kompetenzbereiche auf unterschiedlichen Niveaus anschaulich zu illustrieren. Dabei wurde in Kauf genommen, dass der gewählte Ansatz "the more behavioural view of proficiency" widerspiegelt (North, 2000: 54).

5. Operationalisierung von Sprachkompetenzmodellen

5.1 Im schwarzen Loch des Sprachentestens

Die beiden *State-of-the-art*-Artikel von Bachman (2000) und Alderson & Banerjee (2002) vermitteln beide ein sehr ähnliches Bild: In den rund 20 Jahren seit Beginn der Überlegungen zum "kommunikativen Testen" (vgl. für die Anfänge Morrow, 1979; Canale & Swain, 1980) ist nicht nur in praktischer, sondern auch in theoretischer Hinsicht enorm viel wichtige Arbeit geleistet worden, insbesondere was das Problembewusstsein und das forschungsmethodische und testpraktische Instrumentarium angeht. Aus dem im vorigen Kapitel geschilderten Defizit an gesichertem Wissen in Bezug auf

[15] In der Untersuchung, in der hauptsächlich die Kompetenzbeschreibungen kalibriert wurden, standen die mündlichen Fertigkeiten im Zentrum. Dementsprechend sind die Sprachaktivitäten im Zusammenhang mit dem Lesen und Schreiben eher spärlich beschrieben. Die Sprachmittlung, die Sprachverwendung zugunsten von Dritten, z.B. beim Übersetzen und Dolmetschen, aber auch beim vereinfachenden Vermitteln innerhalb einer Sprache, wurde nicht expliziert. In Glaboniat *et al.* (2005) liegt für Deutsch ein Versuch vor, diese Lücke zu schliessen.

Sprachkompetenzmodelle, auf das Funktionieren von bestimmten Aufgaben in konkreten Kontexten und auf die individuellen Prozesse, die beim Sprachgebrauch und entsprechend auch beim kommunikativen Testen tatsächlich ablaufen, besteht aber aus der Sicht des *language testing* ein Interpretationsproblem. Alderson und Banerjee (2002: 100f.) diagnostizieren gar ein "schwarzes Loch", was die aktuellen Möglichkeiten angeht, Testergebnisse zuverlässig zu interpretieren:

> What we simply do not know at present is […] how trait [d.h. die individuellen kognitiven Dispositionen] and context interact under what circumstances and thus how best to combine the two perspectives in test design. […] Strategies, and presumably traits, can vary across persons and tasks, even when the same scores are achieved. The same test score may represent different abilities, or different combinations of abilities, or different interactions between traits and contexts, and it is currently impossible to say exactly what a score might mean. This we might term The Black Hole of language testing. (Alderson & Banerjee, 2000: 100f.)

Es ist offensichtlich, dass aus einem solchen Diagnose umfassende Forschungsprogramme abgeleitet werden müss(t)en. Fest steht aber auch, dass Testentwickler nicht auf die Ergebnisse der entsprechenden Untersuchungen warten können, sondern, mit der gegebenen Vorsicht, weiterhandeln müssen. Sie sind aber aufgerufen, sich um Validität und Validierung zu bemühen im Rahmen dessen, was aufgrund der Bedeutung der jeweiligen Tests nötig ist (Alderson & Banerjee, 2002: 102).

Testentwickler im Bildungsbereich berufen sich hinsichtlich ihres Validitätsverständnisses häufig auf Messick, der (Konstrukt-)Validität als einen umfassenden Begriff sieht:

> Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment […]. Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. (Messick, 1996: 9)[16]

Charakteristisch für dieses Verständnis von Validität ist, dass es einerseits um die Qualität von Messungen (mit bestimmten Instrumenten bei bestimmten Kandidaten in bestimmten Kontexten) geht und andererseits um die Richtigkeit, Nützlichkeit und Fairness der Interpretation der Ergebnisse (für eine detaillierte Auflistung der einzelnen Punkte siehe Messick, 1996: 10; 12ff.). Zentral wichtig sind nach Messick die Verhinderung von Konstrukt-

---

[16] Einflussreich im Bereich des *language testing* sind wiederum Bachman und Palmer mit ihrem Qualitätsbegriff der *test usefulness* (Bachman & Palmer, 1996: 17ff.). *Test usefulness* wird dabei als Konglomerat gesehen, das verschiedene Qualitätsaspekte von Tests in sich vereint, die in anderen Konzeptionen zum Teil als Antagonismen interpretiert wurden: Reliabilität, Konstruktvalidität, Authentizität, Interaktivität (werden die gewünschten Kompetenzen angesprochen und aktiviert?), *Impact* und Praktikabilität. Die beiden letzten sind testexterne Qualitätskriterien, welche die Auswirken auf das Sprachenlehren und -lernen (*Impact*) bzw. die Grenzen der praktischen Machbarkeit betreffen.

Unterrepräsentation und von Konstrukt-irrelevanter Varianz. Konstrukt-Unterrepräsentation liegt dann vor, wenn ein Test das Konstrukt (die Kompetenzinterpretation; die latenten Könnens- oder Wissensbereiche, über die etwas ausgesagt werden soll) nicht "abdeckt", weil zum Beispiel die Items einseitig ausgewählt wurden oder nicht in genügender Anzahl im Test vorhanden waren. Solchermassen lückenhafte Daten erlauben keine Aussagen über das Konstrukt insgesamt. Konstrukt-irrelevante Varianz ergibt sich zum Beispiel dann, wenn ein Testformat ein bestimmtes Geschick verlangt, das in der Population ungleich verteilt ist und das nichts mit dem zu messenden Konstrukt zu tun hat; komplizierte Computermanipulationen in einem Sprachtest würden vermutlich zu unerwünschter Varianz führen. Weiter kann konstruktirrelevante Varianz auch durch nicht-objektive Formen der Bewertung und Beurteilung entstehen, zum Beispiel in mündlichen Prüfungen.

Messick fordert, dass mit Hilfe von empirischen Daten Belege für die verschiedenen Aspekte von Validität erbracht werden sollen. "Forschendes Testen" sollte also die Regel sein. Dabei ist nicht a priori festgelegt, wie viel Evidenz nötig ist; als Faustregel kann aber gelten, dass der Aufwand, der für die Validierung getrieben wird, der Bedeutung einer Prüfung (vor allem der Konsequenzen daraus) angemessen sein sollte (vgl. Alderson & Banerjee, 2002: 102).

Die Einschränkungen, die weiter oben in Bezug auf ein Kompetenzmodell der kommunikativen Sprachverwendung und auf die Möglichkeit, Testergebnisse zu interpretieren, gemacht wurden, wirken sich natürlich auch auf die Möglichkeiten aus, Validität umfassend zu beurteilen und zu belegen: Mehr als der *State of the art* erlaubt*,* kann nicht verlangt werden[17].

5.2 Konstruktdefinitionen

Eine wichtige Voraussetzung für jede Validierung ist eine genaue Fassung des Testkonstrukts, also dessen, worüber etwas ausgesagt werden soll, zum Beispiel Wortschatzkenntnisse; Kompetenzen im Lesen juristischer Texte usw. Das Konstrukt kann auf unterschiedlichen Ebenen lokalisiert sein. Chapelle (1998, 33ff.) unterscheidet drei theoretische Möglichkeiten der Konstruktdefinition: eine *Trait*-Perspektive, eine behavioristische und eine interaktionistische Perspektive.

Trait-Theoretiker interpretieren Konsistenzen im Test (zufällige Einzelergebnisse sind irrelevant) als Merkmale der Kandidaten. Folglich definieren sie Konstrukte in Form von Kompetenzen und fundamentalen

---

[17] Für Richtlinien bieten sich die einschlägigen Werke der Testliteratur an. In Alderson & Banerjee (2002: 104) ist ein Raster von Luoma abgedruckt, welcher die bei der Entwicklung und Verwendung von Sprachtests zu beachtenden Punkte in Anlehnung an Messick aufführt.

Prozessen. Im Test müssen diese *traits* mit geeigneten Aufgaben angezapft werden; in der Aufgabe und bei der Testdurchführung wird versucht, mögliche Störungsquellen zu eliminieren (z.B. emotional belastende Themen), damit sich das *trait* ungehindert zum Beispiel von negativen Affekten zeigen kann.

Behavioristen schreiben Leistungskonsistenzen im Test Kontextfaktoren (z.B. der Sprecherkonstellation) zu und definieren Konstrukte deshalb mittels der Merkmale des Kontexts. Im Test insgesamt versucht man, eine Serie von Aufgaben bearbeiten zu lassen, die möglichst viele Merkmale mit dem relevanten realweltlichen Sprachgebrauch gemeinsam haben. Als Grundlage für die Testentwicklung werden Analysen der realen Kommunikationsbedürfnisse sowie der realweltlichen Aufgaben durchgeführt (*needs analysis; task analysis*).

Interaktionisten machen gewissermassen eine Synthese der beiden übrigen Ansätze; sie betrachten Konsistenzen als das Ergebnis von *traits,* Merkmalen des Kontexts und der Wechselwirkungen von *traits* mit solchen Merkmalen.

> When trait and context definitions are included in one definition, the quality of each changes. Trait components can no longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics. From the interactionalist perspective, *performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts*. (Chapelle, 1998: 43).

Unter der interaktionistischen Perspektive müssen Test und Aufgaben so konstruiert sein, dass sie die gewünschten Interaktionen zwischen *trait* und Kontext auslösen. Wenn dieses Validitätskriterium erfüllt ist, können einzelne Leistungen als Indizien (*sample*) für die Leistungsfähigkeit in ähnlichen Kontexten genommen werden. Wird der (erleichternde oder erschwerende) Einfluss des Kontexts gebührend berücksichtigt, können Leistungen auch als Zeichen (*signs*) für die latenten Sprachkompetenzen (*traits*) interpretiert werden.

Voraussetzungen für valide Tests und Aufgaben sind einerseits solide theoretische Kompetenzvorstellungen und andererseits Bedürfnis- und Aufgabenanalysen ähnlich denen unter dem behavioristischen Paradigma, allerdings mit geringeren Ansprüchen hinsichtlich einer authentischen Abbildung der realweltlichen Bedingungen; die konkreten Kontextfaktoren treten konzeptuell sozusagen in dasselbe Glied zurück wie die latenten Kandidatenmerkmale (*traits*).

5.3 Generalisierbarkeit

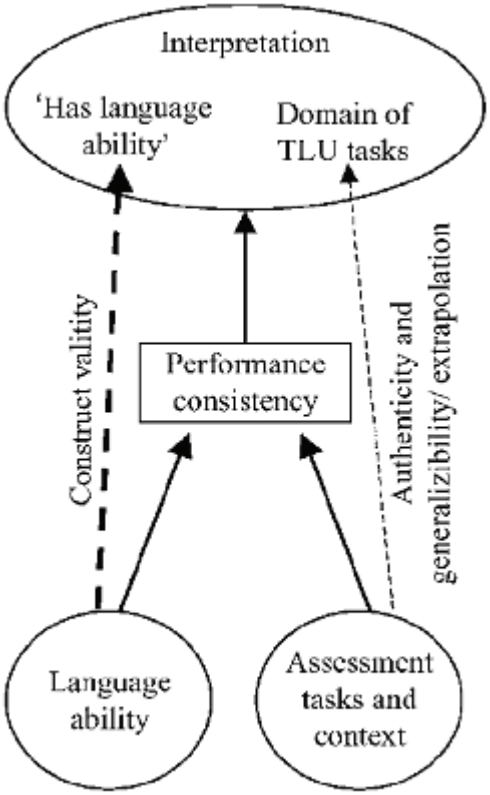


Abb. 6 Interpretation von Testergebnissen nach Bachman (2002: 457)

Die Generalisierbarkeit von Ergebnissen ist ein offensichtliches Grundanliegen von Sprachtests: Sie wären kaum zu rechtfertigen, wenn sie bloss beanspruchten, zu überprüfen, wie gut Kandidaten in Bezug auf genau diesen Test sind, und darüber hinaus keine Aussagen machen wollten, vorzugsweise über "realweltlich" relevante Fähigkeiten. Nach Bachman (2002) interessiert eine Interpretation von Testergebnissen (als *Performance consistency* bezeichnet, weil nur *Konsistenzen* in der Performanz Verallgemeinerungen erlauben) in zweierlei Hinsicht: erstens als Aussage über das Vorhandensein/Ausmass von *language ability* (Fähigkeit zur kommunikativen Sprachverwendung) und zweitens als Aussage über den Wirklichkeitsbereich (*Domain of target-language-use tasks*), in dem erfolgreich Aufgaben bewältigt werden können. Damit diese zweifache Interpretation oder Generalisierung legitim ist, muss zum einen das Konstrukt, d.h. die Modellbildung in Bezug auf *language ability*, valide sein. Weiter muss der Test insgesamt das "Konstrukt" durch die Testaufgaben und Beurteilungskriterien hinreichend repräsentieren, d.h. überprüfbar machen und das Beobachten von *consistencies* überhaupt erlauben (keine Konstrukt-Unterrepräsentation, vgl. Messick, 1994: 14). Zum anderen müssen die Testaufgaben und der Testkontext zu einem gewissen Grad "authentisch" sein, d.h. die relevanten Merkmale der "realen" Kommunikationsaufgaben angemessen repräsentieren, und es dürfen keine

unerwünschten (und unkontrollierten) Test-Umstände und Aufgaben-eigenschaften die Ergebnisse wesentlich beeinflussen (d.h. konstruktir-relevante Varianz verursachen, vgl. Messick, 1994: 14).

Nach Bachman & Palmer (1996: 12) muss bei der Testkonstruktion aufgezeigt werden, dass die Testaufgaben erstens die Kompetenzen der Kandidaten in einer Art mobilisieren, wie sie auch in der entsprechenden "realen" Sprachverwendungsaufgabe mobilisiert würden, und dass Testaufgaben und Testsituation zweitens Merkmale aufweisen, die den Merkmalen der Aufgaben in den "realweltlichen" Aufgaben und Situationen (in relevanten Hinsichten) entsprechen. Nur unter diesen Voraussetzungen dürfe von der Performanz im Test auf den Sprachgebrauch ausserhalb der Testsituation geschlossen werden.

Bachmans Modell der "zweifachen Anbindung" von Konstrukt und Testaufgaben hilft Extreme im Testdesign zu vermeiden: Ansätze zum Testdesign, die primär von der Analyse von Kommunikationsbedürfnissen und von "real" zu bewältigenden Aufgaben ausgehen (*task-driven*), werden dazu angehalten, neben reinen, möglicherweise nur wenig generalisierbaren Performanzgesichtspunkten auch Kompetenzgesichtspunkte in Betracht zu ziehen und Aussagen darüber zu machen[18]; Ansätze, die in erster Linie von Kompetenzvorstellungen im Sinne von mehr oder weniger stabilen Ressourcen oder gar Eigenschaften der Lernenden ausgehen (*construct-driven*, v.a. unter einer *Trait*-Perspektive), sind in Bachmans Modell aufgefordert, den Handlungsbereich und die Handlungsaufgaben zu bestimmen, auf die sich ein Test beziehen soll; dies wiederum hat Konsequenzen für Art und Auswahl der Testaufgaben, für die Beurteilungskriterien sowie insbesondere auch für das Konstruktverständnis.

6. Testentwicklung und -validierung

Wie eingangs erläutert, soll im Rahmen von HarmoS nicht einfach ein Test zur Überprüfung von Bildungsstandards entwickelt werden, sondern es soll ein Kompetenzmodell empirisch überprüft und es sollen Instrumente, insbesondere Testaufgaben mit Problemlösungscharakter, entwickelt und validiert werden. Die beiden Aufgaben unterscheiden sich aber nicht grundsätzlich, denn ein Testen ohne Belege dafür, was wirklich getestet wird, ist ebenso unzulänglich wie eine Modellentwicklung ohne valide Instrumente,

---

[18] Bachman (2002) wendet sich damit in erster Linie gegen das "Task-based language performance assessment" (Norris, Brown, Hudson & Yoshioka, 1998; Norris, 2002; Brown & Hudson, 2002)

mit denen das Modell angemessen operationalisiert und empirisch überprüft werden kann[19].

6.1 Untersuchungs- und Validierungsdesign

Ein vielversprechendes Untersuchungs- und Validierungsmodell, das auch im Bereich des Sprachentestens immer wieder ins Spiel gebracht wird (Mislevy *et al.*, 2002; Bachman, 2000; 2002), scheint das *Evidence-Centred Design* (ECD) aus dem Kreis um Robert Mislevy zu sein (Mislevy, Steinberg, Almond & Russell, 2002; Mislevy, Almond & Lukas, 2004)[20]. Im Folgenden soll aber nicht das ECD selbst dargestellt werden, sondern ein erweitertes Modell, das insbesondere die Verbindung zwischen Sprachverwendung im Test und Sprachverwendung ausserhalb der Testsituation (*target-language use*, um mit Bachman zu sprechen) sowie die Vielfalt der Faktoren in einem interaktionistischen Modell klarer illustriert.

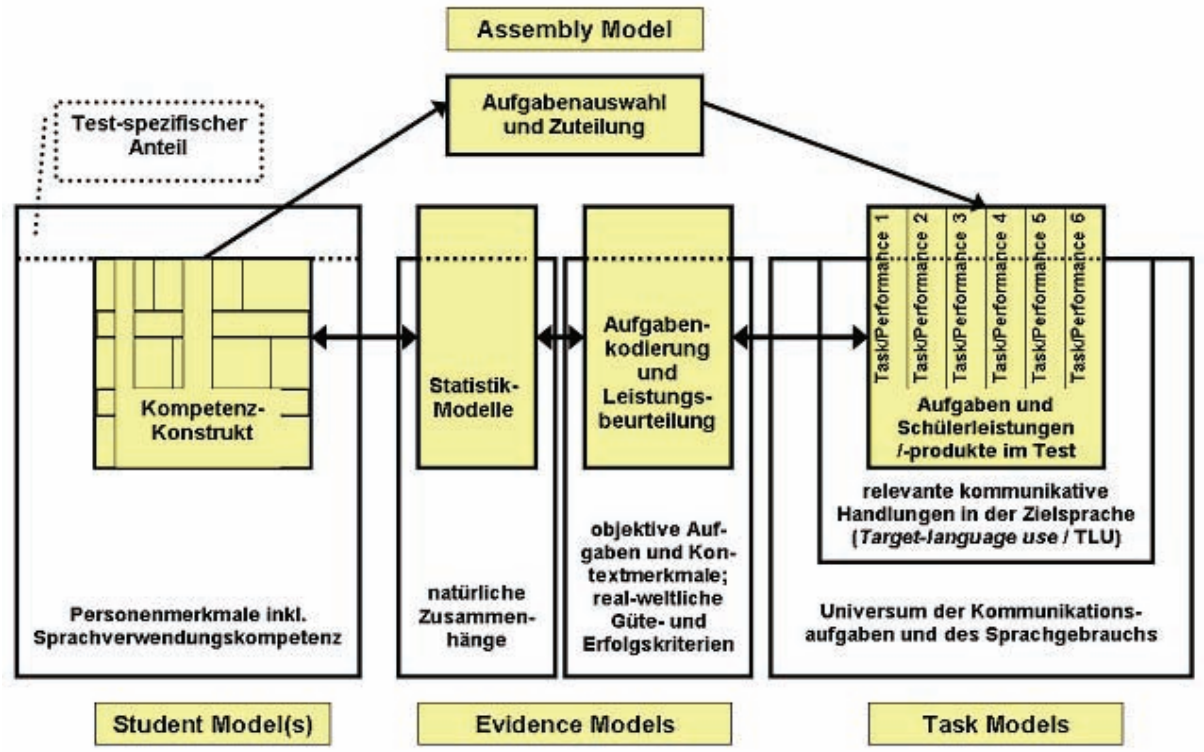


Abb. 7 Erweitertes Modell eines Untersuchungs- und Validierungsdesigns, aufbauend auf das *Evidence-centred design* von Mislevy *et al.* (Vorschlag des Autors, PL)

---

19 Vgl. Alderson & Banerjee (2002: 80) zu den Parallelen zwischen angewandt-linguistischer Forschung und Testentwicklung.

20 Ebenfalls interessant scheint das Modell der *Four building blocks* von Mark Wilson (2004), das dem ECD sehr ähnlich ist und wie dieses das Potential anspruchsvoller Statistikmodelle als konstitutives Element miteinbezieht.

Im (Kompetenz-)Konstrukt des *student model* werden die relevanten Variablen (Wissen und/oder Können) als Kompetenzaspekte spezifiziert. Die Auswahl dieser Aspekte ist oft von externen Überlegungen bestimmt (Was will/muss man über die Population wissen?). Bei den Variablen des *student model* handelt es sich um latente Variablen:

> Operationally, SM [student model] variables summarize patterns of values of observable variables, along lines we build into the evidence model structures […], as evoked by features we build into tasks. (Mislevy *et al*., 2002: 481)

In einem interaktionistischen Modell müssen neben den fokussierten Aspekten des Konstrukts noch weitere Merkmale der Person (kognitive oder Persönlichkeitsfaktoren) in das Design eingeschlossen werden, sofern sie potenziell (und evtl. nur auf gewissen Kompetenzstufen, nicht aber auf anderen) relevant sind. Die empirisch ermittelten Untersuchungsergebnisse bzw. die quantitativen und qualitativen Grenzen eines Untersuchungsdesign zwingen gewöhnlich im Verlauf oder im Anschluss an Untersuchungen zu Modifikationen am Konstrukt; die empirische Validierung eines Konstrukts ist in der Regel ein iterativer Prozess (Hypothesen – Voruntersuchungen – weitere Untersuchungen).

Das Verhältnis zwischen Kompetenzen, Aufgabeneigenschaften und beobachtbaren Variablen erklären Mislevy *et al.* wie folgt:

> Distinct SM variables are used to maintain belief about distinct knowledge/skills, and a claim is associated with particular patterns across them as they are called upon in tasks that stress competences in different ways. Thus, students' proficiency in a domain can be described in terms of which skills they possess (via SM variables), tasks can be described in terms of which skills they require (via TM [task model] variables) and the outcomes expected from any given match-up can be described in terms of values of observable variables. (Mislevy *et al.*, 2002: 484)

In einem konstruktgesteuerten (*construct-driven*) Testentwicklungsansatz kommen die *task models,* also der Test insgesamt mit allen Aufgaben(formen), dadurch zustande, dass man versucht, in genügender Zahl und in der nötigen Vielfalt Gelegenheiten zu schaffen, bei denen die Kompetenzen, die im Konstrukt fokussiert werden, eingesetzt werden müssen. In einem aufgabengesteuerten (*task-driven*) Ansatz würden aufgrund von Analysen des ausgewählten Ziel-Handlungskontexts und der entsprechenden Aufgaben Testaufgaben konstruiert, welche die entsprechenden "realen" Aufgaben repräsentieren. In einem zweiseitigen Entwicklungsansatz, wie er von Bachman gefordert wird (siehe weiter oben), müssen Aufgaben entwickelt werden, die einerseits relevante Aspekte des Ziel-Sprachgebrauchs (TLU) aufnehmen und deren Ergebnisse andererseits genügend Gelegenheit geben, Muster zu beobachten, die Rückschlüsse auf die Kompetenzaspekte des Konstrukts erlauben. Zum Beispiel ist also innerhalb eines konstruktgesteuerten Ansatzes ein dekontextualisierter Grammatiktest denkbar um die grammatische Kompetenz zu messen; innerhalb eines zweiseitigen Ansatzes müsste darauf geachtet werden, dass nicht nur das

Konstrukt "grammatische Kompetenz" in genügender Breite umgesetzt würde (wie im konstruktgesteuerten Ansatz), sondern dass die grammatischen Phänomene auch noch unter unterschiedlichen, relevanten Kontext- und Aufgabenbedingungen in den Aufgaben vorkämen. Ein rein aufgaben-gesteuerter Ansatz würde sich vermutlich nicht um das Konstrukt "grammatische Kompetenz" für sich genommen kümmern, weil kaum kommunikative Aufgaben denkbar sind, die korrekte und/oder komplexe Grammatikverwendung zum eigentlichen Ziel haben. Unter dem Aspekt der Generalisierung bzw. Interpretation von Ergebnissen kann es aber durchaus sinnvoll sein, über rezeptive und produktive grammatische Kompetenz Aussagen zu machen, weil grammatische Kompetenz in praktisch allen kommunikativen Handlungen eine mehr oder weniger wichtige Resource ist.

Die *evidence models* haben die doppelte Funktion, einerseits die relevanten Merkmale der Aufgaben und Lernerleistungen zu erfassen und in Aussagen über die im Konstrukt fokussierten Kompetenzen zu überführen und andererseits zu illustrieren (in Form von erwartbaren Testergebnissen oder produktiven Leistungen mit bestimmten Merkmalen), wie Leistungen aussehen, die bestimmten Ausprägungen des Konstrukts in bestimmten Kontexten entsprechen.

Die Kodierung bezieht sich auf unterschiedliche Aspekte von Aufgaben und Leistungen:

a. auf (vermutlich) konstruktrelevante Merkmale von Input-Texten (Länge, Komplexität usw.), Aufgabenstellungen (Komplexität der Operation, interpersonelle Konstellation usw.) und Produkten (inhaltliche Richtigkeit und Angemessenheit, sprachliche Merkmale usw.);

b. auf (vermutlich) konstruktirrelevante Aspekte wie die Merkmale der Testmethode (z.B. das Aufgabenformat) und Eigenheiten der Beurteilung von Lernerleistungen oder-produkten (v.a. die Beurteilerstrenge und -konsistenz).

Weir (2005a) gibt im Rahmen seiner Überlegungen zur Validierung von Sprachtests eine Reihe von Hinweisen auf Merkmale von Tests und deren Durchführung, welche – in erwünschter oder unerwünschter Weise – die Leistungen beeinflussen können. Die folgende Tabelle zeigt eine Reihe von Faktoren, welche die Ergebnisse von Prüfungen zum Sprechen in erwünschter oder unerwünschter Weise beeinflussen können.

<table>
<tr>
<td rowspan="2">Merkmale der Aufgabe:
· Zweck der Aufgabenstellung
· Antwortformat, Erwartungen an Produkt
· Zeit, Länge
· Personenkonstellation(en)
· Sprachliche Merkmale:
- lexikalisch
- strukturell
- funktional (z.B. Dialogtypen)
· Vorausgesetztes Weltwissen</td>
<td rowspan="2">Interlokutor(en):
· Anzahl
· Sprechgeschwindigkeit
· Akzent(e)
· Grad der Vertrautheit
· Geschlecht</td>
<td>Durchführungsbestimmungen:
· Räumlichkeiten
· Konkrete Durchführung (Vorbereitung usw.)</td>
</tr>
<tr>
<td>Beurteilung:
· Beurteiler (Konsistenz, Strenge)
· Beurteilungskriterien</td>
</tr>
</table>

Abb. 8 Mögliche Quellen von konstruktrelevanter und konstruktirrelevanter Varianz in Tests zum Sprechen (nach Weir, 2005b: 46)

Unter idealen Bedingungen würden alle konstruktrelevanten und konstruktirrelevanten Aufgabenmerkmale kodiert. Unter realen Bedingungen werden vor allem diejenigen Merkmale erfasst, von denen man durch Vorerprobungen und andere Untersuchungen weiss, dass sie eine wesentliche Rolle spielen. Starke konstruktirrelevante Aufgabenmerkmale müssen vermieden bzw. kontrolliert werden (verzerrende Aufgabenformate, Lärm, (zufällige) Unfreundlichkeit von Interlokutoren, nicht reliabel zu handhabende Beurteilungskriterien) da sie sonst verunmöglichen, dass relevante Merkmale von Personen angemessen zur Geltung kommen können.

Bachman (2002: 468) weist darauf hin, dass bei der Kodierung von Aufgaben- und Inputtextmerkmalen besonders darauf geachtet werden muss, ob objektive Gesichtspunkte oder subjektive Einschätzungen erfasst werden; denn wenn zum Beispiel die Schwierigkeit bestimmter Teilaspekte eingeschätzt wird und nachher gezeigt werden kann, dass diese mit der Gesamtschwierigkeit von Aufgaben korrelieren, sei dies im Wesentlichen ein Effekt von Autokorrelation und weniger ein Hinweis auf bestehende Kausalitäten. Wird also zum Beispiel der Wortschatz in einem Text durch Experten aufgrund von Erfahrungen mit Schüler/innen als einfach oder schwierig eingestuft, basieren Korrelationen zwischen einer solchen Einschätzung und der empirisch ermittelten Itemschwierigkeit (auch) auf Autokorrelation. Eine Codierung aufgrund der durchschnittlichen Häufigkeit der vorkommenden Wörter in einem Korpus würde dagegen nicht zu Autokorrelation führen.

Die Auswahl der Beurteilungskriterien für Lernerprodukte oder -leistungen kann sich an unterschiedlichen Wertesystemen orientieren: Real-weltliche Kriterien orientieren sich oft daran, ob eine Aufgabe in funktionaler Hinsicht hinreichend erfüllt ist. Das will allerdings nicht heissen, dass zum Beispiel qualitative Aspekte wie Höflichkeit, Korrektheit oder Flüssigkeit keine Rolle spielen würden; deren Gewichtung müsste aber der Funktion des Anlasses und dem sozialen Kontext entsprechend differenziert erfolgen. Eher

schulische Kriterien orientieren sich daran, worauf im jeweiligen Bildungskontext besonders Wert gelegt wird: vielleicht Korrektheit, Ideenreichtum oder Ästhetik der Darstellung. Wichtig ist, dass Kriterien bewusst ausgewählt werden und dass Gesichtspunkte aus unterschiedlichen Traditionen einander gegenübergestellt und mit rationalen Argumenten gegeneinander abgewogen werden.

Für Autoren wie Bachman oder Mislevy ist die Wahl geeigneter – vor allem komplexer – Statistikmodelle von zentraler Bedeutung. Zum aktuellen Standardverfahren gehören die Bestimmung der empirischen Item-schwierigkeit mittels des Rasch-Modells der Item-response-Theorie und ergänzende Faktoren- und Regressionsanalysen zur Interpretation der empirischen Itemschwierigkeit mithilfe von Aufgaben- und Kontextfaktoren (vgl. z.B. Hartig, J. [im Druck] mit seinen Erläuterungen zum deutschen Grossprojekt *DESI*). Das häufig verwendete *Rasch*-Modell ermöglicht im Wesentlichen den Bau von Skalen einerseits der Itemschwierigkeit und andererseits der Personenfähigkeit, die sich direkt aufeinander beziehen (oft grafisch dargestellt in Form einer so genannten *Wright map*). Da das Modell probabilistisch ist, lässt sich aus dem Skalenwert einer Person prinzipiell die Wahrscheinlichkeit ableiten, mit welcher diese bestimmte Aufgaben auf derselben Skala lösen kann. Die Rasch-Analyse und -Skalierung ist in verschiedener Hinsicht ein sehr praktisches Verfahren, hat aber für sich genommen nur einen beschränkten Nutzen, was Aussagen über Kompetenzmodelle angeht. Bachman (2002: 470) empfiehlt den Einsatz von *Structural equation modeling* um die Wechselwirkungen der verschiedenen Faktoren in einem interaktionistischen Modell aufzuklären, Mislevy *et al.* (2002: 487) erachten unter anderem den Einsatz von Modellen künstlicher Intelligenz (*Bays nets*) als sinnvoll. Vielversprechend erscheint auch der Einsatz von explanatorischen Item-response-Modellen, weil diese die komplexen Interaktionen sowohl auf den Personen als auch auf der Aufgabenseite offenbar angemessener zu modellieren vermögen (vgl. De Boeck & Wilson, 2004).

Das *assembly model* schliesslich legt fest, welche Lernenden anhand welcher Aufgaben und unter welchen Gesichtspunkten beurteilt werden, damit mit optimalem Aufwand die Informationen gewonnen werden können, die nötig sind, um die angestrebten Generalisierungen über Lernende zu ermöglichen, sei es in Form von Aussagen über vorhandene (latente) Kompetenzen oder in Form von Prognosen darüber, welche Klassen von Kommunikationsaufgaben bewältigt werden können. Das *assembly model* erhält eine wichtige Funktion insbesondere in Computer-basierten Test-Designs, wo die Wahl der Aufgaben laufend angepasst werden kann.

7. Zum Untersuchungsdesign in HarmoS

Wie eingangs dieses Artikels erwähnt, hat das Fremdsprachenkonsortium in der *HarmoS*-Hauptuntersuchung von April-Mai 2007 relativ wenig Testzeit zur Verfügung. Im Fertigkeitsbereich Schreiben werden insgesamt (d.h. in den beiden Landesteilen und für drei Zielsprachen) neun verschiedene Aufgabenstellungen eingesetzt, in den Leseverstehenstests für die Landessprachen insgesamt 22 Aufgabenstellungen, d.h. Text(e) und dazugehörige Items; für Englisch sind es nur acht Aufgabenstellungen, weil die Sechstklässler wegfallen. Aus ergänzenden quantitativen und qualitativen Untersuchungen werden weitere Ergebnisse anfallen, besonders zum Sprechen, aber auch zum Hörverstehen und zum Leseverstehen, was qualitative Erkenntnisse angeht.

Weiter ist zu bedenken, dass auf Seiten der Kantone offenbar die Intention besteht, nach der Umsetzung der neuen Szenarien für den Fremdsprachenunterricht eine umfassendere Hauptuntersuchung durchzuführen. In Zukunft ist zudem damit zu rechnen, dass im Zuge des einsetzenden Bildungsmonitorings immer wieder grössere empirische Untersuchungen anfallen werden. Deshalb erscheint es durchaus sinnvoll, zum jetzigen Zeitpunkt umfassende grundsätzliche Überlegungen anzustellen und auch Analysen ins Auge zu fassen, für welche die im Frühjahr 2007 zu erwartende Datenbasis mit Sicherheit zu schmal sein wird.

Im Folgenden soll in Anlehnung an das oben dargestellte Modell kurz skizziert werden, welche Entscheidungen für die Aufgaben- und Testentwicklung hinsichtlich der schriftlichen *HarmoS*-Hauptuntersuchung getroffen wurden:

Zum *student model*:

- Das Gesamtkonstrukt "Kommunikationsfähigkeit in einer Fremdsprache (Deutsch, Französisch oder Englisch)" wird grob nach dem Fertigkeitenschema aufgeteilt: Es sollen in erster Linie Aussagen gemacht werden über die Leseverstehenskompetenz und über die kommunikative Schreibkompetenz. Auf Untersuchungen speziell zur interkulturellen Kompetenz oder zu einzelnen Kompetenzaspekten wie Motivation wird zumindest in der Hauptuntersuchung verzichtet.
- In Bezug auf das Lesekonstrukt sollen speziell Aussagen gemacht werden dazu, wie gut die Schüler/innen (lokal) Informationen entnehmen, Textpassagen (global) interpretieren und Gelesenes zu vorhandenem Wissen in Beziehung setzen (d.h. reflektieren) können. Diese Unterteilung orientiert sich an den kognitiven Prozessen, die beim *Reading-literacy*-Konstrukt von PISA empirisch unterschieden werden konnten. Sie widerspiegelt auch grob die wichtigsten Lesehaltungen, die gemeinhin in der Fremdsprachendidaktik unterschieden werden.

- Beim Schreibkonstrukt werden Makrofunktionen[21], die "aus einer (manchmal längeren) Reihe von Sätzen bestehen" (Europarat, 2001: 126) in den Vordergrund gerückt: Informieren/Beschreiben, Erzählen/-Berichten, Meinung ausdrücken/Argumentieren und Auffordern/-Veranlassen. In einer ersten Näherung wird davon ausgegangen, dass die mit Schrägstrich getrennten Funktionen jeweils ähnliche Kompetenzen mobilisieren. Wichtig erscheint uns an dieser Kategorisierung nach Makrofunktionen, deren grundsätzlich funktionale Orientierung, mit der zentrale Gesichtspunkte eines handlungsorientierung Ansatzes (vgl. *Referenzrahmen*) aufgenommen werden. Kategorisierungen wie die Makrofunktionen gewählte werden in anderen Zusammenhängen auch für eine Kategorisierung von Textsorten oder Diskurstypen nach ihren Hauptfunktionen, dominanten Funktionen oder Globalzielen[22] verwendet. In der Regel bestehen (längere) Texte aus mehreren funktionalen Teilen, von denen oft einer eine übergreifende, dominante Funktion hat. In den Texten, die zu den HarmoS-Schreibaufgaben entstehen, kann mit solchen Schachtelungen gerechnet werden, sodass aus den einzelnen Aufgaben Informationen zu mehr als einer Makrofunktion gewonnen werden können.

Zum *task model*:

- Die Aufgaben orientieren sich grob am Handlungsrahmen und an den Themen, die durch die Deskriptorensammlung des Projekts *IEF/lingualevel* (vgl. Lenz & Studer, 2004; 2005) der Deutschschweizer Kantone umrissen werden; *IEF/lingualevel* baut auf den Beschreibungen von Sprachaktivitäten im *Referenzrahmen* auf, versucht aber, an die Erfahrungswelt der jugendlichen Schüler/innen anzuschliessen[23]. In

---

21 Makrofunktionen sind auch bekannt unter den Bezeichnungen "rhetorische Funktionen" oder "Kommunikationsverfahren".

22 Vgl. Engel, 1988: 118f. Die in *HarmoS* unterschiedenen Funktionen lassen sich relativ leicht Engels "Globalzielen" Informieren, Überzeugen und Veranlassen zuordnen.

23 Zur Verwendung des *Referenzrahmens*, insbesondere der Kompetenzbeschreibungen, als Grundlage für die Testentwicklung ist in letzter Zeit einiges, vor allem auch Kritisches, geschrieben worden. Weir fasst wesentliche Kritikpunkte zusammen: Die Skalen würden nicht genügend Kontextvariablen berücksichtigen und die Bedingungen, unter denen Kommunikationsaufgaben erfolgreich gelöst werden können, nicht genügend beschreiben; zudem kämen sie in den verschiedenen Beschreibungen ungleich vor; Unterschiede in der kognitiven Verarbeitung auf unterschiedlichen Kompetenzniveaus seien kaum berücksichtigt; Handlungsbeschreibungen seien zu selten bezogen auf Beschreibungen der Qualität, die bei der Bewältigung der Aufgaben erwartet wird; der Wortlaut von Deskriptoren sei gelegentlich nicht konsistent oder transparent genug, um daraus Vorgaben für Tests ableiten zu können (Weir, 2005b: 282). Diese Kritik scheint mir im Wesentlichen zutreffend, jedoch ist anzumerken, dass die Skalen nicht in erster Linie als Ausgangsmaterial für Testautoren entwickelt wurden. Im "Manual"-Projekt des Europarats werden seit 2002 Grundlagen geschaffen für die Zuordnung von Prüfungen zu den Referenzniveaus. In diesem Zusammenhang sind bisher

HarmoS wird auch auf Kompatibilität mit den gängigen Fremdsprachenlehrwerken geachtet, was allerdings nicht heisst, dass nur Leistungen abgetestet würden, die genau so auch vorbereitet worden wären. Verschiedentlich wird versucht, Aufgaben durch explizite Situierungen in den (fiktiven) Kontext einer Austauschpädagogik oder persönlicher interkultureller Sprachkontakte zu stellen. Bei den *Input*-Texten zum Leseverstehen wurden kontinuierliche und diskontinuierliche Texte gezielt variiert, auch dies in Anlehnung an die PISA-Untersuchungen zur *reading literacy*. Die letztlich eher kleine Zahl von Aufgaben setzt aber dem Bestreben, Kommunikationskontexte, kommunikative Handlungen und Texttypen breit und unter gezielter Variation von Variablen zu repräsentieren, enge Limiten, sodass es zum Teil nur darum gehen kann, zum jetzigen Zeitpunkt im Hinblick auf spätere Entwicklungen vorzuspuren.

- In den Aufgabenstellungen wird der Handlungsaspekt insofern ernst genommen, als immer eine klare Situierung des Kontexts und eine Motivierung der Aufgabe versucht werden. In den Schreibaufgaben sollen relativ feste inhaltliche Vorgaben den Auftrag bzw. die eigene Intention ersetzen, die ausserhalb von schulischen Situationen in der Regel mit Schreibanlässen einhergehen. Die Aufgaben zu den Texten im Leseverstehen setzen nicht nur die Vorgaben durch das Konstrukt um (Informationsentnahme etc.), sondern wollen gleichzeitig auch ausserschulisch übliche Lesehaltungen und -interessen im Zusammenhang mit den gewählten Texten widerspiegeln.
- Eine Folge aus verschiedenen Entscheidungen zum Untersuchungsdesign ist die Verwendung der lokalen Schulsprache (L1) für Aufgabenstellung und -beschreibung, und auch als Sprache der Schüler/innen bei offenen Antworten im Leseverstehen. Dazu ein paar Überlegungen im Einzelnen:
- Das Schreiben soll getrennt von der Lesefertigkeit getestet werden; bei den Leseverstehenstests soll klar sein, was getestet wird (nämlich Lesekompetenz in Bezug auf die Input-Texte, nicht die Itemformulierung, jedenfalls im Rahmen von herkömmlichen Aufgabenformaten wie *multiple-choice* oder richtig-falsch).

---

beim Europarat zwei nützliche Publikationen (Council of Europe, 2003; 2004) entstanden; weiteres Material für die Zuordnung und Entwicklung von Prüfungen ist in verschiedenen Teilprojekten, die aus dem "Manual"-Projekt hervorgegangen sind, entstanden oder noch im Entstehen (vgl. www.coe.int/portfolio > "CEFR and related documents").

- Beim Lesen soll "Verstehen" auf der Bedeutungsebene getestet werden, nicht nur auf der Ebene der Erkennung von Wort*formen,* wie dies bei rein zielsprachlich gehaltenen Tests oft der Fall ist.
- Beim Schreiben soll der Wortschatz zur Lösung der Aufgaben nicht im Einführungstext relativ unkontrolliert vorgegeben werden. Ausserdem soll den Schüler/innen Aufgabenstellung und situierung klar sein, denn beim handlungsorientierten Testen ist eine Aufgabenstellung mehr als ein blosser Prompt im Stil von "Ecris une histoire intéressante!"

Zum *evidence model:*

Die konkreten Arbeiten am *evidence model* sind zum gegenwärtigen Zeitpunkt noch nicht sehr weit gediehen. Bei der Entwicklung der Leseaufgaben wurden versuchsweise erste Kodierungen von konstruktrelevanten (bezogen auf *student* und *task model*) und konstruktirrelevanten Merkmalen vorgenommen[24]. Zudem wurden vor der Finalisierung praktisch alle Schreib- und Leseaufgaben in möglichst unterschiedlichen Klassen erprobt, und es wurden auch Feedbacks von Lehrpersonen und anderen Fachleuten eingeholt. Die Ergebnisse der Erprobung wurden sehr sorgfältig ausgewertet, die Ergebnisse zu den Leseverstehensaufgaben auch mithilfe von statistischen Methoden. Dagegen war es unter dem enormen Zeitdruck, der aufgrund der am Anfang dieses Beitrags erwähnten organisatorischen Probleme bestand, nicht möglich, alle (insbesondere die rezeptiven) Aufgaben rechtzeitig vor dem Druck auch mit qualitativen Methoden der angewandten Linguistik wie zum Beispiel dem *stimulated recall protocol* zu validieren (vgl. Banerjee, 2004); von den qualitativen Untersuchungen wären vor allem Hinweise darauf zu erwarten gewesen, welche Kompetenzen bei der Bearbeitung der Aufgaben denn tatsächlich involviert sind – und ob es sich um konstruktrelevante oder vorwiegend irrelevante Kompetenzen (Geschicktheit im Umgang mit bestimmten Aufgabenformaten usw.) handelt. Es ist vorgesehen, die qualitativen Untersuchungen vor der Auswertung der Hauptuntersuchung zumindest teilweise nachzuholen, um die Interpretation der Ergebnisse auf eine möglichst sichere Basis zu stellen. Die Schülertexte, die bei der Erprobung der Schreibaufgaben entstanden, werden in nächster Zeit dazu verwendet, die ersten umfassenderen Kodierungsschemata zu entwerfen. Bei den Beurteilungskriterien zum Schreiben wird es in erster Linie darum gehen, die nicht-einzelsprach- und einzelaufgabenspezifisch gefassten Kriterien aus *IEF/lingualevel* (und z.T. auch aus dem *Referenzrahmen*) zu konkretisieren und damit möglichst objektiv anwendbar zu machen, bevor im Frühjahr aussenstehende Beurteiler/innen eingeführt werden müssen.

---

[24] Ein Ausgangspunkt sind auch in diesem Punkt die Untersuchungen im Rahmen von *DESI* (vgl. Nold & Willenberg (im Druck): 23-41) sowie Alderson *et al.*, (2004).

Die statistische Auswertung der Hauptuntersuchung wird eine bereits bestehende Gruppe von Statistikern und Methodologen besorgen. Sie werden hauptsächlich die Software *Conquest* einsetzen, die auch bei PISA Verwendung findet. *Conquest* bietet mehr als einfache *Rasch*-Programme, indem es zum Beispiel auch Mehrdimensionalität in den Daten darstellen oder neben Itemschwierigkeit und Schülerkompetenz weitere so genannte Facetten wie Beurteilerkonsistenz und -strenge schätzen kann. Zur Gewinnung von vertieften Einsichten hinsichtlich eines Kompetenzmodells werden ergänzend auch herkömmliche Methoden der Varianzaufklärung eingesetzt.

Zum *assembly model:*

Eine wichtige Fragestellung bei den Aufgabenvorerprobungen war jene nach der Akzeptanz und Lösbarkeit der verschiedenen Aufgaben in der sechsten und in der neunten Klasse. Da es ein Ziel ist, aufgrund der Untersuchungsergebnisse möglichst eine kontinuierliche Kompetenzskala zu bilden, ist es notwendig, dass eine Reihe von Aufgaben von beiden Populationen bearbeitet wird.

Insgesamt werden die Sechst- und Neuntklässler/innen in der Deutsch- und Westschweiz in einem Matrixdesign getestet: Ein einzelner Schüler wird sich in der Hauptuntersuchung während maximal 70 Minuten mit einer Fremdsprache beschäftigen, die meisten weniger lange. Dafür werden alle Aufgaben (bzw. Testhefte) auf zufällige Stichproben von jeweils 150 Schüler/innen verteilt, sodass es möglich ist, die Leistungsmittelwerte der verschiedenen Gruppen gleichzusetzen und für alle Test-Items eine einzige Skala zu bilden.

Dadurch dass die einzelnen Schüler/innen anlässlich der Hauptuntersuchung in verschiedenen Fertigkeiten und Sprachen (L1 plus eine der beiden Fremdsprachen) getestet werden, wird es möglich sein verschiedene Korrelationen zu überprüfen – sowohl zwischen Fertigkeiten innerhalb einer Sprache als auch über Sprachen hinweg. So dürfte es beispielsweise interessant sein, die Lese- oder Schreibkompetenz in der L1 mit der entsprechenden Kompetenz in einer Fremdsprache zu vergleichen oder natürlich die Lese- und Schreibkompetenz in einer Fremdsprache. Speziell soll auch die Beziehung zwischen den Ergebnissen im C-Test mit den Ergebnissen im Schreiben untersucht werden, weil da mit interessanten Übereinstimmungen zu rechnen ist.

8. Schlussbemerkung

Es war mit Sicherheit ein mutiger Schritt der EDK, die Initiative zu ergreifen für ein Entwicklungsprodukt, wie es jetzt im Zusammenhang mit den Vorarbeiten für Bildungsstandards in vier Fächern bzw. Fächerbereichen entstanden ist. Das Wagnis und die Anforderungen für die Fachkonsortien sind nicht klein.

Das ganze Unterfangen wird sich aber dann lohnen, wenn mehr entsteht, als nur gedruckte Bildungsstandards und ein Reservoir an Testaufgaben, das gelegentlich zwecks Bildungsmonitoring nachgefüllt wird. Zur Zeit werden Kompetenzen aufgebaut und Grundlagen entwickelt, für die es schlicht schade wäre, wenn sie nicht durch die eine oder andere Form der Institutionalisierung kontinuierlich weiterentwickelt würden.

BIBLIOGRAPHIE

Alderson, J.Ch. & Banerjee, J. (2002): Language testing and assessment (part 2). State-of-the-art review. In: Language Teaching, 2, 79-113.

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004): The Development of Specifications for Item Development and Classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening. Online: www.ling.lancs. ac.uk/cefgrid. (26.01.2007)

Bachman, L.F. (1990): Fundamental considerations in language testing. Oxford (Oxford University Press).

Bachman, L.F. (2004): Statistical analyses for language assessment. Cambridge (Cambridge University Press), 2004.

Bachman, L.F. (2002): Some reflections on task-based language performance assessment. In: Language Testing 19, (3), 453-476.

Bachman, L.F. (2000): Modern language testing at the turn of the century: assuring that what we count counts. In: Language Testing 17, (1) 2000, 1-42.

Bachman, L.F. & Palmer, A.S. (1996): Language testing in practice: designing and developing useful language tests. Oxford (Oxford University Press).

Banerjee, J. (2004): Qualitative analysis methods. In: Council of Europe (Hg.): Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment, Section D. Strasbourg (Council of Europe) (DGIV/EDU/LANG (2004) 13). Online: www.coe.int/portfolio > Documentation. (26.01.2007)

Brown, J.D. & Hudson, Th. (2002): Criterion-referenced language testing. Cambridge (Cambridge University Press).

Bygate, M., Skehan, P. & Swain, M. (2001): Researching pedagogic task. Second Language Learning, Teaching and Testing. Harlow (Pearson Education).

Canale, M. & Swain, M. (1980): Theoretical bases of communicative approaches to second language teaching and testing. In: Applied Linguistics 1 (1), 1-47.

Chapelle, C. (1998): Construct definition and validity inquiry in SLA research. In: Bachman, L. & Cohen, A. (Hg.): Interfaces between second language acquisition and language testing research. Cambridge (CUP), 32-70.

Chapelle, C., Grabe, W. & Berns, M. (1997): Communicative language proficiency: definitions and implications for TOEFL 2000. Monograph Series. Princeton (N.J.: ETS).

Council of Europe (2003). Relating language examinations to the Common European framework of reference for languages (CEF). Manual, Preliminary pilot version. Strasbourg (Language Policies). Online: http://www.coe.int/T/E/Cultural_Cooperation/education/Languages/-Language_Policy/Manual/. (26.01.2007)

Council of Europe (2004). Relating language examinations to the Common European framework of reference for languages (CEF). Reference supplement to the preliminary pilot version of the manual. Strasbourg (Language Policies). Online: http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/9srefsupl.asp. (26.01.2007)

Davies, A. (1989): Communicative competence as language use. In: Applied Linguistics, 10 (2), 157-170.

De Boeck, P. & Wilson, M. (Hg.) (2004): Explanatory item response models: A generalized linear and nonlinear approach. New York (Springer).

Douglas, D. (2000): Assessing language for specific purposes: theory and practice. Cambridge (Cambridge University Press).

EDK/Schweizerische Konferenz der kantonalen Erziehungsdirektoren (2004): Sprachenunterricht in der obligatorischen Schule: Strategie der EDK und Arbeitsplan für die gesamtschweizerische Koordination. Bern. Online: http://www.edk.ch/PDF_Downloads/Presse/REF_B_31-03-2004_-d.pdf. (26.01.2007)

Engel, U. (1988): Deutsche Grammatik (2. Aufl.). Heidelberg (Julius Groos).

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

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H. & Wertenschlag, L. (2005): Profile deutsch. Version 2.0. Berlin (Langenscheidt).

Hartig, J. (im Druck): Skalierung und Kompetenzniveaus. In: [Projektbericht zu DES/], 83-99.

Hymes, D.H. (1972): On communicative competence. In: Pride, J. & Holmes, J.: Sociolinguistics, 269-291.

Klieme, E. *et al.* (2003). Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Hrsg. Bundesministerium für Bildung und Forschung. Bonn (BMBF). Online: http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf. (26.01.2007)

Lenz, P. & Studer, T. (2004): Sprachkompetenzen von Jugendlichen einschätzbar machen. In: Babylonia 2/2004, 21-25.

Lenz, P. & Studer, T. (2005): Neue Instrumente für die Beurteilung der Französisch- und Englischkompetenzen von Deutschschweizer Schülerinnen und Schülern. In: Gohard-Radenkovic, A. (Hg.): Plurilingualisme, interculturalité et didactique des langues étrangères dans un contexte bilingue / Mehrsprachigkeit, Interkulturalität und Fremdsprachendidaktik in einem zweisprachigen Kontext. Transversales, 11. Bern u.a. (Peter Lang), 37-56.

Luoma, S. (2004): Assessing speaking. Cambridge (Cambridge University Press).

McKay, P. (2005): Assessing young learners. Cambridge (CUP).

McNamara, T. (1996): Measuring second language performance. New York (Longman).

Messick, S. (1989): Validity. In: Linn, R.L. (Hg.) Educational measurement. 3. Aufl. New York (American Council on Education/Macmillan), 13-103.

Messick, S. (1994): The interplay of evidence and consequences in the validation of performance assessments. In: Educational Researcher 23 (2), 13-23.

Messick, S. (1996): Validity and washback in language testing. Princeton (ETS Research Report).

Mislevy, R.J., Almond, R.G. & Lukas, J.F. (2004): A brief introduction to evidence-centered design. CSE Report 632. Los Angeles (National Center for Research on Evaluation).

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2002): Design and analysis in task-based language assessment. In: Language Testing, 477-496.

Morrow, K. (1979): Communicative language testing: revolution or evolution? In: Brumfit, C.J. & Johnson, K. (Hg.): The communicative approach to language teaching. Oxford (Oxford University Press), 143-57.

Nold, G, & Willenberg, H. (im Druck): Lesefähigkeit. In: [Projektbericht zu DES/], 23-41.

Norris, J.M. (2002): Interpretations, intended uses and designs in task-based language assessment. In: Language Testing, 337-346.

Norris, J.M., Brown, J.D., Hudson, T. & Yoshioka, J. (1998): Designing second language performance assessments. Honolulu (University of Hawaii Press).

North, B. (2000). The development of a common framework scale of language proficiency. New York (Peter Lang).

North, B. & Schneider, G. (1998). Scaling Descriptors for Language Proficiency Scales. In: Language Testing 15, 2, 217-262.

NW EDK (Hg.) (2007): lingualevel. Instrumente zur Evaluation von Fremdsprachenkompetenzen. 5.-9. Schuljahr. Bern (Schulverlag).

OECD (Hg.) (2003): The PISA 2003 assessment framework. Online: http://www.oecd.org/dataoecd/46/14/33694881.pdf. (26.01.2007)

Robinson, P. (Hg.) (2001): Cognition and second language instruction. Cambridge (Cambridge University Press).

Robinson, P. (2005): Cognitive complexity and task sequencing: studies in a componential framework for second language task design. In: IRAL 43 (1), 1-32.

Schneider, G. & North, B. (1999). "In anderen Sprachen kann ich..." – Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit. Umsetzungsbericht. Bern, Aarau (Nationales Forschungsprogramm 33) (NFP33), Schweizerische Koordinationsstelle für Bildungsforschung (SKBF).

Schneider, G. & North, B. (2000a). "Dans d'autres langues je suis capable de ..." Echelles pour la description, l'évaluation et l'auto-évaluation des compétences en langues étrangères. Rapport de valorisation. Berne, Aarau (Programme national de recherche 33) (PNR33), Centre suisse de coordination pour la recherche en éducation (CSRE).

Schneider, G. & North, B. (2000b). Fremdsprachen können – was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit. Chur, Zürich (Rüegger).

Skehan, P. (1998): A cognitive approach to language learning. Oxford (Oxford University Press).

Skehan, P. (2001): Tasks and language performance assessment. In: Bygate, M., Skehan, P. & Swain, M. (Hg.): Researching pedagogic tasks: second language learning, teaching and testing. New York (Pearson Education), 167-85.

Spolsky, B. (1989): Knowledge of language and ability for use. In: Applied Linguistics, 10 (2), 138-156.

Weir, C. (2005a): Language testing and validation. An evidence-based approach. London (Palgrave Macmillan).

Weir, C. (2005b): Limitations of the Common European framework for developing comparable examinations and tests. In: Language Testing, 3, 281-300.

Wilson, M. (2004): Constructing measures. An item response modeling approach. Mahwah, N.J (Erlbaum).

Widdowson, H.G. (1989): Knowledge of language and ability for use. In: Applied Linguistics, 10 (2), 128-137.