



<https://bulletin.vals-asla.ch> · e-ISSN: 2813-7973 · p-ISSN: 1023-2044

Studer, T. (2004). Internationale Zertifikate für Deutsch als Fremdsprache in der Schweiz. *Bulletin suisse de linguistique appliquée*, 79, 69-97. <https://doi.org/10.5169/seals-978456>

Dieser Artikel ist unter der *Creative Commons Attribution 4.0 International* Lizenz veröffentlicht (CC BY):
<https://creativecommons.org/licenses/by/4.0>



© Thomas Studer, 2004

Internationale Zertifikate für Deutsch als Fremdsprache in der Schweiz

Thomas STUDER

Lern- und Forschungszentrum Fremdsprachen, Bereich Deutsch als Fremdsprache,
Universität Freiburg/Fribourg, Criblet 13, CH-1700 Fribourg; thomas.studer@unifr.ch

Les certificats modernes sont des tests de langue standardisés favorisant une approche communicative. Leurs rapports transparents avec un continuum de compétence langagière leur confèrent une haute valeur de signification pour toutes les personnes impliquées dans le processus. Ainsi, les certificats ne peuvent pas seulement contribuer à améliorer la comparabilité et la reconnaissance des compétences en langues étrangères, mais ils peuvent également avoir impact positif sur l'enseignement et l'apprentissage des langues étrangères.

Dans cet article, on donnera d'abord un aperçu de la situation actuelle des certificats de l'Allemand langue étrangère, en tenant compte avant tout de la situation en Suisse. Ensuite seront brièvement discutées quelques questions concernant la recherche en test de langue. Après quelques informations sur la coopération trinationale dans le domaine des examens Allemand langue étrangère, on présentera de manière plus détaillée un nouveau projet d'examen également important pour la Suisse: il s'agit du "Zertifikat Deutsch für Jugendliche". Nos réflexions porteront notamment sur la mise en relation de cet examen avec les catégories et le système de niveaux du "Cadre européen commun de référence pour les langues" et sur l'utilisation de cet examen dans des écoles publiques.

Zum Thema

Parallel zum hohen Stellenwert, den Fremdsprachen heute haben, gewinnen Beurteilung und Evaluation von Fremdsprachenkenntnissen allgemein und die Zertifizierung dieser Kenntnisse besonders an Bedeutung (vgl. u.a. Morfeld 2003; Schiffko 2001; Schneider 2000). Entsprechend gross ist das Angebot an allgemeinsprachlichen und berufsbezogenen Fremdsprachenzertifikaten, und dieses Angebot ist in letzter Zeit auch für Deutsch als Fremdsprache (DaF) nochmals deutlich grösser geworden; neu hinzugekommen sind v.a. auch Prüfungen für spezielle Zielgruppen wie Studierende, Kinder und Jugendliche sowie MigrantInnen.

Zertifikate können wesentlich dazu beitragen, die Vergleichbarkeit und Anerkennung von Fremdsprachenkenntnissen zu verbessern. Und genau das – Vergleichbarkeit und Anerkennung – ist nötig und wird immer wichtiger, wenn Fremdsprachen heute unter ganz verschiedenen Bedingungen, zu verschiedenen Zeitpunkten und mit unterschiedlichen Methoden gelernt und erworben werden. Eine wichtige Voraussetzung für Vergleichbarkeit ist Transparenz: Vergleichbar sind sprachliche Leistungen dann, wenn sie klar und verständlich beschrieben und – diese Einsicht setzt sich immer mehr durch – wenn sie auf ein gemeinsames Drittes bezogen, d.h. auf ein Referenzsystem

projiziert werden, wie es mit dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER; deutsche Version: Europarat, 2001) vorliegt. Das ist allerdings nicht von heute auf morgen zu haben, sondern ist ein aufwändiger Prozess. Morfeld (2003, 384) charakterisiert ein Zertifikat als „Zusatzqualifikation“ und sieht den Nutzen eines Fremdsprachenzertifikats hauptsächlich in der „Transparenz der nachgewiesenen Leistung und ihrer Relevanz für Arbeitgeber oder Bildungsinstitutionen“. Das trifft zweifellos den Kern der Sache, nur sind Schlüsselwörter wie ‘Transparenz’, ‘Vergleichbarkeit’ und ‘Relevanz der bescheinigten Leistungen’ mit Blick auf die angedeuteten Projektionen nicht einfache Prädikate von Zertifikaten, sondern eher progressive Forderungen und primär einmal Postulate, die es aufzuweisen gilt.

Heute gibt es kaum mehr ein Zertifikat (und übrigens ja auch kaum mehr ein Lehrwerk), von dem nicht behauptet wird, dass es z.B. auf dem Niveau A1 oder B1 – ja was jetzt genau? Dass das Zertifikat auf dem Niveau B1 ‘liegt’ oder ‘im Bereich von B1 situiert ist’ oder ‘auf dieses Niveau bezogen’ ist? Oder ist das Zertifikat auf eine Bandbreite von (Fein-)Niveaus bezogen und steht eine mit soundso vielen Punkten bestandene Prüfung für das Erreichen des Niveaus? Die im Einzelnen sehr verschieden formulierten Zuordnungen sind indessen noch das kleinere Problem. Das grössere Problem besteht darin, dass Indizien, die diese Zuordnungen stützen könnten, derzeit noch weitgehend fehlen und/oder nicht ausgewiesen werden. Dadurch ist eine Art von Scheinobjektivität entstanden, die nur vordergründig Klarheit schafft, in Wirklichkeit aber zu einer neuen Unübersichtlichkeit führt. So wird paradoxerweise genau das wieder in Frage gestellt, was die Zuordnung von Prüfungen zu einem gemeinsamen Referenzsystem primär leisten soll: Vergleichbarkeit.

Vor diesem Hintergrund soll im Folgenden ein aktueller Überblick über Zertifikate für DaF gegeben werden, wobei besonders die Verhältnisse in der Schweiz fokussiert werden (Kap. 1). Ein kurzer und subjektiver Blick auf Fragen der Sprachtestforschung ist Gegenstand von Kap. 2. Im 3. Kap. wird die trinationale Kooperation im Prüfungsbereich DaF etwas näher vorgestellt, an der die Schweiz beteiligt ist. Im letzten Abschnitt (Kap. 4) wird über ein neues, auch für die Schweiz relevantes Prüfungsprojekt berichtet: das Zertifikat Deutsch für Jugendliche. Einen Schwerpunkt bilden dabei Überlegungen zur Zuordnung dieser Prüfung zu den Kategorien und zum Niveausystem des GER.

1. Überblick über Zertifikate für DaF

In Abbildung 1 sind die derzeit verfügbaren Zertifikate für DaF tabellarisch aufgelistet. Fett gedruckt sind diejenigen Zertifikate, die auch in der Schweiz eine grössere Rolle spielen.

*Niveau	**Prüfung	Zielgruppe	***Trägerschaft
A1	Start Deutsch 1	Erwachsene	GI; WBT
	Start Deutsch 1 z	Zuwanderer in Deutschland	GI; WBT
	Sprachkenntnisnachweis Deutsch (SKN)	Zuwanderer in Österreich	ÖSD
	Fit in Deutsch 1	Kinder und Jugendliche in Italien und Frankreich	GI
	Kompetenz in Deutsch 1 (KID 1)	Kinder und Jugendliche in mehreren europ. Ländern	ÖSD
A2	Start Deutsch 2	Erwachsene	GI; WBT
	Grundbaustein Deutsch (ab 1.1.05 durch Start Deutsch 2 ersetzt)	Erwachsene	WBT
	Grundstufe Deutsch	Erwachsene	ÖSD
	Start Deutsch 2 z	Zuwanderer in Deutschland	GI; WBT
	Fit in Deutsch 2	Kinder und Jugendliche in Italien und Frankreich	GI
	Kompetenz in Deutsch 2 (KID 2)	Kinder und Jugendliche in mehreren europ. Ländern	ÖSD
A2+	Sprachstandsanalyse Deutsch (SaD)	'Aussiedler' in Deutschland	WBT
	Test Deutsch	Einbürgerungswillige in mehreren dt. Bundesländern	WBT
	<i>Test Arbeitsplatz Deutsch (TAD)</i>	Erwachsene	WBT
B1	Zertifikat Deutsch (ZD)	Erwachsene	GI; WBT; ÖSD; EDK
	Zertifikat Deutsch für Jugendliche (ZD j; ab 4. Quartal 2004)	Jugendliche (12-16jährige)	GI; WBT; ÖSD; EDK
B2	Zertifikat Deutsch plus (ZD plus)	Erwachsene	WBT
	Zertifikat Deutsch für den Beruf (ZDfB)	Erwachsene	GI; WBT
C1	Zentrale Mittelstufenprüfung (ZMP)	Erwachsene	GI
	Mittelstufe Deutsch	Erwachsene	ÖSD
	Prüfung Wirtschaftsdeutsch International (PWD)	Erwachsene	GI

C2	Zentrale Oberstufenprüfung (ZOP)	Erwachsene	GI
	Kleines Deutsches Sprachdiplom (KDS)	Erwachsene	GI
	<i>Wirtschaftssprache Deutsch</i>	Erwachsene	ÖSD
	Grosses Deutsches Sprachdiplom (GDS)	Erwachsene	GI

Abb. 1: Zertifikate für DaF¹

- * Niveau: Gemäss GER; in der Regel Selbstaussagen der Prüfungsanbieter (oft Expertenkonsens).
- ** Notation der Prüfungen: normale Schrift=allgemeinsprachliche Prüfung; *kursiv*=berufsbezogene Prüfung; **fett**=Prüfung, die auch in der Schweiz angeboten wird.
- *** Trägerschaft: GI=Goethe-Institut; WBT=Weiterbildungs-Testsysteme GmbH; ÖSD=Österreichisches Sprachdiplom Deutsch; EDK=Schweizerische Konferenz der kantonalen Erziehungsdirektoren.

Zwei Bemerkungen zu Abbildung 1:

- Die Tabelle zeigt sog. Niveauprüfungen. Nicht aufgeführt sind computer-adaptive Einstufungstests wie *Dialang* und andere diagnostisch ausgerichtete Prüfungen wie *Bulats* oder *TestDaF*, die mehrere Niveaus umfassen und die deshalb mit der gewählten Darstellungsart nicht adäquat erfasst werden können². Nicht aufgeführt sind ferner auch Fremdsprachen-Zertifika­tionssysteme wie *UN/cert*, die sich als Rahmenvorgaben für Prüfungen verstehen (vgl. Voss, 2001).
- Die in der Tabelle genannten Trägerorganisationen informieren im Internet ausführlich und anschaulich über die von ihnen angebotenen Prüfungen, vgl. <http://www.goethe.de/dll/prf/deindex.htm> (Goethe-Institut, München); <http://www.sprachenzertifikate.de> (WBT Weiterbildungstestsysteme GmbH, Frankfurt a.M.); <http://www.osd.at> (Österreichisches Sprachdiplom Deutsch, Wien). Ausserdem kann man sich auf diesen Websites auch über das recht unterschiedliche Profil der drei Institutionen informieren.

1 Bei der Tabelle handelt es sich um eine aktualisierte und modifizierte Version einer Zusammenstellung, die sich auf der Schweizer Portfolio Website findet und die ihrerseits auf die Publikation "Vivere le lingue" (Dipartimento dell'istruzione e della cultura del Cantone Ticino, ed., 2001) zurückgeht (vgl. http://www.sprachenportfolio.ch/esp_d/hintergrund/index.htm).

2 *Dialang* (www.dialang.org) ist ein vollständig Web-basiertes Diagnosesystem, das im Rahmen eines sehr breit angelegten EU-Projekts entwickelt wurde. Das System ist frei verfügbar, hat Benutzerführungen in 14 Sprachen und erlaubt eine differenzierte Diagnose der Sprachkompetenzen in denselben 14 Sprachen. Die erreichten Fremdsprachenkompetenzen werden durch eine Kombination von Selbstbeurteilungsschecklisten und kalibrierten Testaufgaben diagnostiziert, die auf den Kompetenzbeschreibungen des GER aufbauen. *Dialang* gibt ein unmittelbares Feedback in Form von Angaben zum jeweils erreichten Niveau und von Empfehlungen zum Weiterlernen; es werden aber keine Zertifikate ausgestellt. *Bulats* (Business Language Testing Service, <http://www.bulats.org>) ist ein umfassender, teilweise computer-adaptiver Einstufungstest für 4 Sprachen im beruflichen Kontext. *TestDaF* (Test Deutsch als Fremdsprache, <http://www.testdaf.de>) wendet sich vor allem an Studierende, die ein Studium in Deutschland aufnehmen möchten und die hierfür einen Nachweis ihrer deutschen Sprachkenntnisse auf fortgeschrittenem Niveau brauchen.

Dazu gehört z.B., dass DaF-Prüfungen für das GI als weltweit tätiges Kulturinstitut der BRD nur ein Teil der Arbeit sind, dass die WBT unter dem Namen TELC (The European Language Certificates) Prüfungen für insgesamt 11 Sprachen anbietet und dass das ÖSD auch eine wissenschaftliche Leitung auf universitärer Ebene hat.

1.1. *DaF-Zertifikate in der Schweiz*

In der Schweiz sind das GI und die WBT durch Partnerorganisationen vertreten; das ÖSD bietet derzeit in der Schweiz keine Prüfungen an. Schweizer Lizenznehmer des GI gibt es in Basel, Biel, Freiburg, Genf, Lausanne, Lugano, Neuenburg, Sitten und Zürich; Lizenznehmer sind Universitäten (Université de Lausanne, Università della Svizzera Italiana, Lugano), Hochschulen (Hochschule Wallis, Zürcher Hochschule Winterthur) und verschiedene Weiterbildungsinstitutionen (z.B. das Institut de formation des adultes, Genève). Prüfungspartner der WBT in der Schweiz sind: der Verband der Schweizerischen Volkshochschulen (Bern), der Migros-Genossenschaftsbund (Koordinationsstelle der Klubschulen, Zürich) und die Flying Teachers (Zürich).

Wie aus Abbildung 1 hervorgeht, werden in der Schweiz momentan 8 Zertifikate für DaF angeboten. Was das Volumen dieser Prüfungen betrifft, so können hier dank der Auskünfte von Frau Gabriele Gauler (GI-Zentralverwaltung, München) und Herrn Heinrich Rübeling (WBT, Frankfurt a.M.) Angaben "aus erster Hand" gemacht werden: In der Schweiz werden jährlich insgesamt ca. 7000 DaF-Prüfungen abgelegt; etwa je die Hälfte davon wird von den Partnerorganisationen des GI bzw. der WBT durchgeführt. Beim GI wurde die *ZMP* besonders stark nachgefragt, gefolgt vom *ZD*; anteilmässig weniger bedeutend waren demgegenüber das *ZDfB* und das *GDS* (Angaben für 2003). Bei der WBT wurde das *ZD* weitaus am stärksten nachgefragt.

Neben diesen Prüfungen werden in der Schweiz seit kurzem auch *TestDaF* (an der Hochschule Wallis) und neuerdings auch *Start Deutsch* (vorerst nur an der Universität Lugano) und, mit sehr kleinem Volumen, das *ZD plus* angeboten. Dass es auch in der Schweiz einen Bedarf gibt für Prüfungen auf den unteren Niveaus und dass deshalb die *Start-Prüfungen* Einzug halten, steht ausser Frage. Dass es aber, von Einzelfällen abgesehen, in der Schweiz ein Publikum für *TestDaF* gibt, ist nicht unmittelbar einsichtig, da es sich ja um eine Prüfung für den Hochschulzugang in Deutschland handelt. Im Bereich der Hochschulzugangs-Prüfungen läuft in der Schweiz jedenfalls seit längerer Zeit

ein Projekt mit dem Ziel, die grosse Heterogenität der Anforderungen auszugleichen, die mit den verschiedenen Prüfungen einhergehen. Darüber hinaus wird versucht, zu einem Testmodell zu kommen, das für die Studiensprachen Deutsch, Französisch und Italienisch gleichermaßen gilt³.

Nicht angeboten werden in der Schweiz die sog. Migrantenprüfungen, die in Deutschland (*Start Deutsch Z* und *Test Deutsch*) und in Österreich (*Sprachkenntnisnachweis Deutsch*) entwickelt wurden. Das ist auch nicht wünschenswert, schon aus Gründen der Sprachsituation in der Deutschschweiz nicht (dazu und zum Schweizerhochdeutschen vgl. Werlen sowie Schmidlin und Bickel in diesem Heft). Wenn ein Sprachtest für Einbürgerungsinteressierte in der Schweiz breiter politischer Konsens wäre, müsste an eine Eigenentwicklung gedacht werden (Fragen dazu stellt auch Studer, 2001), oder aber an eine umfassendere Adaptation z.B. des österreichischen Modells, das besonders durch die Kombination von Verfahren der Selbsteinschätzung und Testaufgaben einerseits und durch ein gut durchdachtes, integrales Konzept von Sprachprüfung und Sprachkurs andererseits zu überzeugen vermag. Wenig überzeugend scheint dagegen der Sprachtest, den die Berner Gemeinde Ostermundigen als Teil des Einbürgerungsverfahrens eingeführt hat. Wenig überzeugend v.a., weil es sich dabei, Presseberichten nach zu urteilen, um einen Grammatik- und Wortschatztest im Paper-and-Pencil-Format handelt, nicht um eine kommunikative Prüfung, die weit bessere Chancen hätte, das Sprachenlernen des Zielpublikums positiv zu beeinflussen (zum sog. "impact" von Prüfungen vgl. unten, Kap. 2.2).

1.2. Charakteristika von Zertifikaten

Bei den in Tabelle 1 aufgeführten Zertifikaten handelt es sich grundsätzlich um *kommunikative Prüfungen*⁴ und damit um eine Generation von Tests, die,

3 Koordinator dieses Projekts ist Peter Sauter, Universität Freiburg/ Fribourg.

4 Die Begriffe 'Zertifikat', 'Prüfung' und 'Test' werden in der Literatur nicht einheitlich gebraucht. Tentativ lässt sich zwar, etwa mit Bezug auf Krause & Sändig (2002, 168ff.) das Folgende sagen: Ein Test ist 'kleiner' als eine Prüfung; mit Tests werden z.B. Teilfertigkeiten der kommunikativen Kompetenz oder psycholinguistische Prozesse überprüft. Eine Prüfung dagegen ist etwas Umfassenders, eine Art komplexer Test, mit dem z.B. verschiedene Teilfertigkeiten der kommunikativen Kompetenz überprüft werden. Entsprechend wäre ein Zertifikat eine Prüfung, zu der es ein mehr oder weniger prestigeträchtiges Zeugnis gibt. Aber das ist nicht einheitlicher Sprachgebrauch.

Die Beschreibungen in diesem Abschnitt basieren, abgesehen von der jeweils erwähnten Literatur, allgemein auch auf dem GER (178ff.) sowie auf Vollmer (2003) und Gogolin (2001).

bezogen auf Spolskys Modell der Entwicklung von Sprachtests im 20. Jahrhundert (Spolsky 1976, zitiert nach Davies 2003, 356f.), der dritten Phase in dieser Entwicklung zugeordnet werden können. Diese dritte und jüngste, auch “psycholinguistisch-soziolinguistisch” genannte Phase wird gegen eine “vorwissenschaftliche” (erste Phase) und von einer “psychometrisch-strukturalistischen” (zweite Phase) abgegrenzt, wobei in der Literatur v.a. auch das Nacheinander dieser Phasen unterschiedlich eingeschätzt wird (vgl. neben Davies auch Vollmer 2003; McNamara 2000; Schneider 2000). Kommunikative Prüfungen zielen auf kommunikatives Handeln, auf die Sprachverwendung in Situationen, nicht auf isolierte Elemente der Sprachfähigkeit wie Aussprache, Wortschatz und Grammatik. In kommunikativen Sprachtests geht es zwar auch darum, Prozesse der Sprachverarbeitung anzustossen und die resultierenden Verstehens-, Sprech- oder Schreibleistungen zu beobachten und zu bewerten (‘psycholinguistische Perspektive’). Gleichzeitig aber wird versucht, die KandidatInnen in kommunikative Akte zu involvieren, d.h. in den Tests werden soziale Rollen definiert und die Aufgaben werden kommunikativ situiert, um die sozialen Sprachfunktionen abzubilden bzw. mitzumodellieren (‘soziolinguistische Perspektive’).

Mindestens tendenziell, d.h. bezogen v.a. auf die Sprech- und Schreibaufgaben, sind die aufgeführten Zertifikate als *Performanztests* anzusehen, nicht als Kompetenztests: Kompetenztests arbeiten mit Modellen ‘tieferliegender’ Fähigkeiten, die das kommunikative Verhalten steuern (sollen), und entsprechend spielen in solchen Tests der Kontext von Items sowie die Kontextualisierung und Teilnehmerorientierung von Aufgaben in der Regel keine oder nur eine marginale Rolle. Im Gegensatz dazu modellieren Performanztests ‘reales’, d.h. als real angesehenes zielsprachliches Verhalten. Natürlich handelt es sich auch bei Performanztests nicht um ‘echte’, sondern um simulierte bzw. inszenierte Kommunikation, die, je nach Qualität der Aufgaben, bessere oder schlechtere Anhaltspunkte dafür liefert, wie sich eine Testperson ausserhalb des Testzusammenhangs verhalten *könnte* (“testing is about making inferencies”, wie es McNamara (2000, 7ff.) ausdrückt).

Etwas anders verhält es sich mit den Aufgaben zum Hören und Lesen: Zwar werden auch die Aufgaben zur Überprüfung rezeptiver Fertigkeiten auf der Basis von Konzepten realen kommunikativen Verhaltens modelliert und entsprechend situiert, aber gearbeitet wird in der Regel, besonders auf tieferen und mittleren Niveaus, mit geschlossenen Antwortformaten wie Richtig-Falsch

oder Mehrfachauswahl. Die Begründung hierfür ist, dass Fertigkeiten wie das Hörverstehen separat – d.h. nicht integrativ, also nicht in Kombination mit anderen Fertigkeiten – getestet und bewertet werden sollen. Produktive Formate zur Überprüfung rezeptiver Kompetenzen wie z.B. das Schreiben einer Zusammenfassung würden, so das Argument, zu Validitätsproblemen führen (Komponenten der mündlichen oder schriftlichen Ausdrucksfähigkeit können das Bild rezeptiver Leistungen stark verfälschen; vgl. z.B. Glaboniat 1998, 128). Eine Folge dieses unterschiedlichen Vorgehens bei der Überprüfung von produktiven und rezeptiven Fertigkeiten ist, dass die resultierenden Daten ungleich 'reich' sind, und unterschiedlich sind entsprechend auch die Anhaltspunkte, welche diese Daten für Aussagen über das kommunikative Verhalten ausserhalb der Testsituation liefern: Vielfältig und reich in dem Sinne, dass z.B. auch strategische Aspekte kommunikativen Verhaltens beobachtbar werden, und gut im Sinne von direkteren Anhaltspunkten für Inferenzen sind in erster Linie die Daten zu den produktiven Fertigkeiten. Vergleichsweise ärmer – und 'dafür' häufig sehr reliabel – sind dagegen die Daten zu den rezeptiven Fertigkeiten.

Zertifikatsprüfungen wie die angeführten sind *proficiency tests* (Leistungs- oder Qualifikationsprüfungen), d.h. es geht um die Überprüfung von Sprachfertigkeiten unabhängig davon, wann, wie und wo diese Fertigkeiten gelernt und/oder erworben wurden. Das unterscheidet Zertifikatsprüfungen von *achievement tests* (Lernfortschritts- oder Sprachstandstests): Lernfortschrittstests wie z.B. Prüfungen in Sprachkursen oder auch in Schulen sind mit spezifischen Lehr-Lern-Kontexten verbunden. Der Hauptzweck solcher Prüfungen besteht darin, festzustellen, ob und wo Lernfortschritte in Bezug auf bestimmte, oft eng eingegrenzte und kleinschrittige Lernziele gemacht wurden. Entsprechend sind Lernfortschrittstests eher rückwärts gerichtet (Was wurde als Resultat eines Lehr-Lernprozesses gelernt?), wohingegen Leistungsprüfungen stärker prospektiven Charakter haben: Es interessiert v.a., was das Prüfungsergebnis über sprachliches Verhalten nach der Testsituation aussagt.

Ein weiterer und bedeutender Unterschied zwischen Zertifikaten und kleineren, z.B. traditionellen schulischen Tests ist, dass erstere *kriterienreferenziert*, letztere dagegen 'normorientiert', d.h. auf eine bestimmte Gruppe von LernerInnen bezogen sind: Kriterienorientierte Bewertung erfolgt, indem individuelle Prüfungsergebnisse auf ein Kontinuum von Sprachkompetenz (bzw. auf ein Modell davon) bezogen werden. Für 'normorientierte' Bewertung ist charakteristisch, dass individuelle Leistungen relativ zu den durch-

schnittlichen Leistungen in einer Gruppe, z.B. relativ zum Klassendurchschnitt, festgestellt werden. Diese Terminologie ist allerdings nicht unproblematisch, denn auch kriterienorientiertes Testen und Bewerten ist nicht normfrei (vgl. dazu grundsätzlich Shohamy 2001 und z.B. Glaboniat 2002, die beim kriterienorientierten Testen mit Recht von einer "Sachnorm" spricht).

Zertifikate gelten oft als *high stakes tests*, obwohl das eigentlich keine Eigenschaft von Zertifikaten, sondern eine Funktion bzw. ein Aspekt des Gebrauchs von Prüfungen ist: Werden Zertifikate im Zusammenhang mit Entscheidungen eingesetzt, die weitergehende Konsequenzen für die TestkandidatInnen haben (z.B. Anstellung, Beförderung, Zulassung zur Universität), kann man ihnen dieses Prädikat zuschreiben.

2. Ein Blick auf die Sprachtestforschung

Aktuellere Überblicke über die Sprachtestforschung bieten u.a. Alderson & Banerjee (2001/2002) und Bachman (2000). Aus der Vielfalt von Fragestellungen, Untersuchungen und Kontroversen im Bereich der Sprachtestforschung seien an dieser Stelle wenigstens zwei Punkte herausgegriffen, die mit Zertifikaten in engerem Zusammenhang stehen. Angesprochen werden sollen einerseits die Problematik von Hypothesen über Sprachkompetenzen, die auf Resultaten von inszenierter Kommunikation basieren (Aspekte der Konstruktvalidität eines Tests, 2.1), und andererseits Fragen, die sich im Zusammenhang mit den Wirkungen und den Konsequenzen von "high stakes tests" stellen ("impact" von Prüfungen, 2.2). Aus Platzgründen lediglich erwähnt sei, dass es seit einigen Jahren auch eine Diskussion um Schnittstellen ("interfaces") zwischen Sprachtestforschung und Zweitspracherwerbsforschung gibt (dazu der Sammelband von Bachman & Cohen 1998), die zeigt, dass und in welchen Bereichen die beiden Disziplinen voneinander profitieren könnten – wenn sie sich denn (vermehrt) zur Kenntnis nehmen würden (vgl. dazu auch Berthele in diesem Heft).

2.1. Konstruktvalidität

Allgemein gesagt bezieht sich der Begriff Konstruktvalidität darauf, inwieweit ein Testresultat als Indikator für die zu messende Fähigkeit oder Fertigkeit interpretiert werden kann. Aber was genau ist das zu Messende und wie lässt sich dieses operationalisieren und so in einen Test umsetzen, dass vom Testresultat gesagt werden kann: "[...] reflects the area(s) of language ability

we want to measure, *and very little else.*”? (Bachman & Palmer 1996, 21; Hervorheb. Stu.)

Zwischen sprachlichen Leistungen, wie sie in Tests sichtbar werden, und sprachlichem Handeln ausserhalb der Testsituation gibt es immer eine Lücke, auch bei Performanztests. Das liegt daran, dass zielsprachliches Verhalten, das sich ja in singulären Kontexten vollzieht, nur als Abstraktion und damit nur als Konstrukt fassbar ist. Aber diese Lücke kann mehr oder weniger gross sein, je nachdem, wie das Konstrukt in einen Test umgesetzt wird. Bei der Umsetzung unterscheidet McNamara (2000, 53) zwei problematische Fälle: Zum einen können TestkandidatInnen durch die Testinhalte ebenso wie durch die Aufgaben- und Antwortformate eines Tests mit Anforderungen konfrontiert werden, die kein ‘Gegenstück’ in ihrem sonstigen Sprachgebrauch haben bzw. die für ihr zielsprachliches Handeln irrelevant sind (“construct irrelevant variance”). Zum andern kann es sein, dass ein Test bestimmte Aspekte der fremdsprachlichen Kommunikation, für die man sich interessiert, nicht oder nur ungenügend repräsentiert (“construct under-representation”).

In der Forschung lassen sich sehr grob zwei Hauptlinien ausmachen: Auf der einen Seite die theoretische Diskussion um das Konstrukt ‘an sich’, also um das Modell von Sprachkompetenz, das einem Test zugrunde liegt, und auf der andern empirische Untersuchungen zu testbezogenen Faktoren, die das Testergebnis beeinflussen und so zu unzulässigen Hypothesen über das zielsprachliche Verhalten führen können. Auf der Seite der Testtheorie gibt es derzeit Kontroversen um das meistzitierte Modell in der Literatur zum Testen und Prüfen, das ist das Modell der ‘Sprachfähigkeit’ (“model of communicative language ability”) von Bachman (1990; zitiert nach Bachman & Palmer 1996, Kap. 4). Die Kontroversen betreffen besonders den Allgemeinheitsanspruch dieses Modells und die Frage, ob es ein psychologisches *und* soziologisches sei/sein könne (vgl. McNamara, 2003 467f.). Im Bereich der empirischen Untersuchungen sind v.a. Forschungen im Umkreis von aufgabenbasierten Test-Konzepten (“task-based approaches to testing”) zu nennen (vgl. Bygate et al. 2001, Teil III, und dort besonders Skehan 2001). Erforscht werden dort Aufgabenparameter, die das Testergebnis beeinflussen und die deshalb Rückschlüsse auf sprachliche Fähigkeiten und Fertigkeiten gefährden können.

2.2. Impact

Sprachtests allgemein und prestigeträchtige Prüfungen wie Zertifikate besonders bzw. die Interpretation und Nutzung von Testergebnissen haben

vielfältige Aus- und Rückwirkungen, primär auf das Lehren und Lernen in Kursen oder im Unterricht (Mikroebene), dann aber auch auf das jeweilige curriculare Umfeld, das Bildungssystem und die Gesellschaft insgesamt (Makroebene), und diese Wirkungen können positiv oder negativ sein (z.B. Davies 2003; Wall 2000; McNamara 2000; Bachman & Palmer 1996)⁵. Schifko (2001, 831) fasst den Forschungsstand zu Rückwirkungsmechanismen in Form von fünf Hypothesen zusammen, mit denen verschiedene Wirkungsbereiche voneinander abgegrenzt werden sollen (im Einzelnen begründet werden diese Hypothesen indessen nicht): Unterschieden werden curriculare Aspekte (Tests haben Auswirkungen auf Lehrpläne und Lernzielkataloge sowie auf die Gestaltung und/oder die Auswahl von Lehrwerken und Unterrichtsmaterialien), didaktisch-methodische und lernstrategische Aspekte (Tests haben Auswirkungen auf das Was und das Wie des Unterrichtens und Lernens) und motivationspsychologische Aspekte (Tests wirken sich auf Einstellungen zu den Inhalten und Methoden von Unterricht und Lernen aus). Quer dazu stehen zwei Hypothesen zur Stärke und zur Breite dieser Effekte: Schifko geht davon aus, dass v.a. auch der Status eines Tests über die Stärke der Effekte entscheidet und dass diese Effekte für bestimmte Lernende und Lehrende gelten, für andere dagegen nicht.

Eine Hauptkontroverse in der aktuellen Forschung zum “impact” dreht sich darum, welche der effektiven und potentiellen Auswirkungen eines Tests in den Verantwortungsbereich von Personen und Institutionen gehören, die Prüfungen entwickeln, anbieten und durchführen (vgl. Davies, 2003, 355).

Fakt ist, dass empirische Untersuchungen zum “impact”, insbesondere solche, welche die Makroebene der Auswirkungen von Tests (mit)betreffen, bis heute sehr selten sind⁶. Hingegen zeichnet sich in der Literatur so etwas wie eine begriffliche ‘Umschichtung’ dieses Phänomens ab (v.a.: “impact”-Aspekte als Komponenten eines übergeordneten Konzepts der Testfairness, vgl. Kunnan 2000).

5 Terminologisch wird “impact” oft (explizit bei Davies 2003, 361) als Oberbegriff für diese Auswirkungen gebraucht, wohingegen “washback” (manchmal auch “backwash”) eher für unmittelbarere Rückwirkungen von Tests auf Lernsituationen steht.

6 Ein erhebliches Defizit ist das besonders auch im Zusammenhang mit den sog. Migrationsprüfungen: Hier bedeutet Mangel an Forschung ganz direkt, dass man jenseits von Teilnehmerzahlen kaum etwas Gesichertes über den Gebrauch von solchen Prüfungen weiss – und damit v.a. auch nichts über einen allfälligen Missbrauch.

3. Mitbestimmen und Mitarbeiten: Die Rolle der Schweiz in der trinationalen Kooperation im Prüfungsbereich DaF

Seit rund 10 Jahren gibt es im Prüfungsbereich DaF eine trinationale Kooperation, in der die folgenden Institutionen vertreten sind: Das GI und die WBT für Deutschland, das ÖSD für Österreich und die EDK für die Schweiz. Zweck der Kooperation ist die gemeinsame Revision, Entwicklung und Fortschreibung internationaler DaF-Zertifikate für Hauptniveaus des GER. Die EDK hat, handelnd auch für die Bundesämter BBW (*Bildung und Wissenschaft*) und BBT (*Berufsbildung und Technologie*), den Bereich DaF des Lern- und Forschungszentrums Fremdsprachen (LeFoZeF) der Universität Fribourg/Freiburg beauftragt, in dieser Kooperation mitzuarbeiten. Die Hauptgründe für die Beteiligung der Schweiz sind:

- DaF-Zertifikate werden auch in allen Landesteilen der Schweiz angeboten und abgelegt (vgl. Kap. 1.1), und sie spielen z.B. im Bildungswesen und auf dem Arbeitsmarkt eine bedeutende Rolle. Folglich liegt es im Interesse der Schweiz, dass diese Zertifikate von hoher Qualität sind, internationalen Standards entsprechen und (daher) europaweit oder sogar weltweit anerkannt werden.
- DaF-Zertifikate stellen für alle beteiligten Akteure und besonders auch für die Lernenden eine Art Modell der Zielsprache und der zielsprachlichen Kommunikation dar, und sie transportieren kulturelle Vorstellungen und Werte der deutschsprachigen Länder. Deshalb hat die Schweiz ein Interesse daran, dass schweizerische Besonderheiten in diesen Zertifikaten mitberücksichtigt werden, und zwar so, dass ein zeitgemässes und adäquates Bild der Schweiz in sprachlicher *und* kultureller Hinsicht vermittelt wird.

Hauptprodukte der Kooperation sind das ZD, zu dem mittlerweile ca. 25 Prüfungssätze entwickelt wurden, und verschiedene zugehörige Materialien wie v.a. die sog. Lernzielbroschüre (LZB; vgl. EDK/ GI/ ÖSD/ WBT, Hrsg., 1999), in welcher Ansatz, Elemente und Prinzipien des ZD im Detail vorgestellt werden und die eine Reihe von Lernzielkatalogen umfasst, darunter auch Wortschatzlisten und eine didaktische Grammatik. Mitgearbeitet hatte die Schweiz (und ebenso Österreich) bereits an der Revision der ZMP. Derzeit stehen die Fortschreibung und Qualitätssicherung des ZD sowie das Projekt ZD j (vgl. Kap. 4) im Vordergrund der Arbeiten.

Die Rechte und Pflichten der Partner in Bezug auf die gemeinsam getragene Prüfung ZD sind in einer Kooperationsvereinbarung festgehalten. Zu den für die Schweiz besonders wichtigen Punkten dieser Vereinbarung gehören:

- Sprachliche und landeskundliche Besonderheiten aller drei beteiligten Länder werden berücksichtigt. Das bedeutet v.a. auch, dass die Plurizentrik als Zertifikatsprinzip garantiert ist. Realisiert wird dieses Prinzip dadurch, dass in jedem Testsatz neben bundesdeutschem und österreichischem Deutsch auch Schweizerhochdeutsch vorkommt.
- Die gemeinsame Trägerschaft der Prüfung ZD ist auf den Zertifikatsurkunden und auf allen aus der Kooperation hervorgehenden Publikationen ersichtlich. Dadurch werden Mitarbeit *und* Mitverantwortung der Schweiz auch nach aussen sichtbar.
- Die Kooperationspartner können Lizenzen für die Abnahme des ZD vergeben.

Der letzte Punkt hat für die Schweiz eine spezielle Bedeutung: Sie könnte zwar Lizenzen vergeben, macht das aber nicht, weil sie (bis jetzt und im Gegensatz zu vielen andern europäischen Ländern) kein eigenes Prüfungszentrum und -vertriebssystem unterhält. Angeboten und durchgeführt wird das ZD in der Schweiz, wie oben dargestellt, von den Schweizer Partnerorganisationen des GI und der WBT.

Die Beteiligung der Schweiz an der Kooperation umfasst die aktive Mitarbeit in der Steuerungs- und in der Redaktionsgruppe ("Test-AG") der Kooperation⁷. In der Steuerungsgruppe werden u.a. Massnahmen zur Qualitätssicherung des ZD diskutiert und getroffen (darunter z.B. kleinere Untersuchungen zur Reliabilität der Bewertungen im Prüfungsteil Schriftlicher Ausdruck) sowie, andererseits, Initiativen für künftige Prüfungsprojekte besprochen. In der Test-AG werden die Testentwürfe für Prüfungs-, Modell- und Übungssätze des ZD – und jetzt auch des ZD j – bearbeitet (vgl. auch Studer 2000).

Die trinationale Kooperation im Prüfungsbereich DaF ist auch in einem weiteren Zusammenhang zu sehen: Zusammenarbeit und Zusammenschlüsse im Bereich der Fremdsprachenzertifikate sind in Europa seit längerem üblich, und sie werden auch als Ausdruck einer weiter zunehmenden Professionalisierung des Prüfungswesens gesehen (vgl. Morfeld 2003; Schifko 2001). Die grösste europäische Organisation ist die *Association of Language Testers* in

7 Michael Langner, Günther Schneider und der Schreibende teilen sich diese Arbeit.

Europe (ALTE)⁸. In dieser Organisation sind derzeit (2004) 28 europäische Prüfungsinstitutionen zusammengeschlossen, die insgesamt 24 Sprachen repräsentieren (vgl. <http://www.alte.org>). Die trinationale Kooperation für DaF ist zwar um viele Schuhnummern kleiner als die ALTE, aber gerade für Deutsch als – weltweit gesehen kleine und, im Vergleich zu früher, weniger bedeutende – Fremdsprache (zu den Zahlen vgl. z.B. Krumm 2003, 40f.) scheint eine Zusammenarbeit der deutschsprachigen Länder aus sprachpolitischer Sicht sehr sinnvoll und zweckdienlich. Für die Schweiz als *auch* deutschsprachiges und zudem sehr kleines Land mit entsprechend beschränkten Ressourcen gilt das in ganz besonderem Mass.

4. Das Zertifikat Deutsch für Jugendliche: Warum und wie eine neue Prüfung entsteht⁹

Das Zertifikat Deutsch für Jugendliche (ZD j) ist eine standardisierte, kommunikativ orientierte Leistungs- oder Qualifikationsprüfung für Deutsch als Fremdsprache, die im Bereich des Niveaus B1 des europäischen Referenzrahmens situiert sein sollte¹⁰. Sie richtet sich an Jugendliche im Alter von 12-16 Jahren und soll sowohl im deutschsprachigen Inland als auch weltweit angeboten werden. Die Prüfung wird von den 4 Institutionen GI, WBT, ÖSD und EDK im Rahmen der vorgestellten Kooperation entwickelt, und zwar in enger Anlehnung an die von den Partnern gemeinsam erarbeiteten Lernziele und Testformate des Zertifikats Deutsch (vgl. EDK et al. 1999). Das ZD j befindet sich in der Endphase der Entwicklung und sollte ab dem 4. Quartal 2004 einsatzbereit sein.

Im Folgenden wird zuerst das sog. "Manual" (Europarat 2003) vorgestellt (4.1) und als Grundlage für eine etwas genauere Beschreibung des ZD j herangezogen (4.2). Anschliessend werden Motive nachgezeichnet, die für die Entwicklung der neuen Prüfung ausschlaggebend waren (4.3), sowie ausgewählte Aspekte der laufenden Entwicklungsarbeiten skizziert (4.4). Abge-

8 Ziel von ALTE ist es, gemeinsame Sprachkompetenzstufen (ursprünglich 5, jetzt 6, wie im GER und kompatibel damit) sowie gemeinsame Standards für alle Phasen der Testentwicklung und -auswertung zu etablieren. Hingearbeitet wird damit auf die länderübergreifende Anerkennung von Zertifikaten und Diplomen – auch, aber sicher nicht nur, aus ökonomischen Interessen.

9 Für spannende Diskussionen und wertvolle Informationen rund um das ZD j bedanke ich mich besonders bei den KollegInnen der Test-AG der Kooperation, namentlich bei Ursula Schmitz (GI), Elisabeth Piskernik und Gerhild Kronberger (ÖSD) sowie Heinrich Rübeling (WBT).

10 Die niveaubezogene Aussage ist – vorderhand – ein Postulat im Sinne der Einleitung zu diesem Artikel, wenn auch eines, das sich auf die LZB und grundsätzlich auf die "Kontaktschwelle" stützen kann (vgl. Baldegger et al. 1980).

geschlossen werden die Ausführungen mit Überlegungen zum Einsatz des ZD j (4.5).

4.1. Zum “Manual”

Mit dem “Manual” des Europarats¹¹ wird derzeit ein Werkzeug entwickelt, das dazu dienen soll, verschiedene Prüfungen und insbesondere auch Zertifikate und Diplome durch die Vermittlung des GER vergleichen zu können. Vorgesprochen werden eine Reihe von Prozeduren und Aktivitäten, die dabei helfen sollen, Ansprüche (“claims”), die mit einer Prüfung in Bezug auf den GER verbunden werden – allgemein: “die Prüfung ist auf den Referenzrahmen bezogen”, speziell z.B.: “die Prüfung liegt im Bereich des Niveaus B1” –, besser abzustützen und transparent zu gestalten (“validation of the claims”). Für die Validierung der Ansprüche werden vier Schritte empfohlen, die aufeinander aufbauen und die zu immer zuverlässigeren Indizien für die postulierten Bezüge führen sollen. Die vier Schritte und die damit verbundenen Hauptziele sind: *Familiarisation* (vertieftes Verständnis des Ansatzes und besonders der Kompetenz- und Niveaubeschreibungen des GER), *Specification* (qualitative Beschreibungen der Prüfung in Kategorien und Begriffen des GER sowie in Bezug auf allgemeinere testtheoretische Standards), *Standardisation* (v.a. auch Schulung von BewerterInnen anhand von standardisierten, den GER-Niveaus zugeordnete Leistungen sowie Erstellen von solchen Referenzleistungen zu den eigenen Prüfungen) und *Empirical Validation* (Datenerhebung und –analyse im Zusammenhang mit eingesetzten Testsätzen, um die Befunde der vorangehenden Schritte empirisch weiter abzustützen).

Das “Manual” ist *eine* konsequente Weiterführung des GER, und als solche ist es zweifellos von grossem Nutzen, wenn es um mehr Transparenz und Vergleichbarkeit im Prüfungswesen geht. Nicht leicht zu fassen ist indessen der genaue Anspruch dieses ‘Handbuchs’, und je nach Perspektive wird man darin auch den einen oder andern kritischeren Aspekt ausmachen. Dabei ist zu bedenken, dass das “Manual” den Status eines Entwurfs hat und weiterentwickelt werden soll. Insofern verstehen sich die drei folgenden Beispiele für kritischere Aspekte des “Manuals” als Verbesserungsvorschläge:

11 Council of Europe (Ed., 2003): Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual. Preliminary Pilot Version, November 2003.

- Einerseits wird explizit gesagt, das “Manual” sei keine Anleitung zur Testkonstruktion (S. 3). Andererseits werden oft geschlossene (!) Tabellen mit Kategorien und Elementen wie z.B. “Typen von Testaufgaben” oder “Aspekte der Validität” (S. 35 u. 37) präsentiert, welche die Testkonstruktion im Kern betreffen und die, eben wegen dieser ‘hermetischen’ Präsentation, den Eindruck erwecken, dass sie für *die* “good practice” stehen. So kommt nicht in den Blick, dass es zu einigen dieser Komponenten eine theoretische Diskussion gibt, die keineswegs abgeschlossen ist (vgl. Kap. 2), und es ist fraglich, ob beim Abarbeiten von Checklisten eine vertiefte Reflexion über die Qualität von Prüfungen in Gang kommt, wie sie vom “Manual” intendiert ist.
- Ein anderer Punkt ist, dass das “Manual” eine ganz starke Prämisse macht bzw. auf einem sehr hohen Niveau einsteigt, wenn vorausgesetzt wird, dass Prüfungsanbieter die Konstruktvalidität einer Prüfung nachweisen, *bevor* sie sich auf die Zuordnung der Prüfung zum GER einlassen (S. 3). Ob diese Prämisse realistisch ist, sei dahin gestellt, denn gerade im ebenso zentralen wie komplexen Bereich der Konstruktvalidität einer Prüfung fehlt es an konkreten Hilfestellungen.
- Nicht unproblematisch scheint auch, dass das Kriterium “externe Validität” nun auch für die Beziehung zwischen einer Prüfung und dem GER verwendet wird (z.B. S. 4). Dieses Kriterium steht sonst in der empirischen Forschung allgemein für das Verhältnis von Stichprobe und Population bzw. dafür, ob ein stichprobenabhängiges Untersuchungsergebnis auf andere Personen, Situationen oder Zeitpunkte generalisiert werden kann (vgl. z.B. Bortz & Döring 1995, 53). Es ist nicht einzusehen, warum dieses Kriterium jetzt auch für die Referenz auf ein Bezugssystem wie den GER verwendet werden sollte.

4.2. Einige Elemente und Prinzipien des ZD j

Für deskriptive Zwecke bietet Schritt zwei des “Manuals” (Kap. 4, “Specification”) eine gute Grundlage. Präsentiert werden dort insgesamt 23 Formulare, mit denen sich Prüfungen beschreiben lassen. Die Formulare decken ein breites Spektrum von Facetten einer Prüfung ab und reichen von den Prüfungskordinaten über Aspekte der Prüfungsentwicklung, -analyse und -überarbeitung, der Bewertung von Prüfungsleistungen und der Mitteilung von Prüfungsergebnissen bis hin zu einer detaillierten Auseinandersetzung mit Fragen, die sich beim Überprüfen einzelner kommunikativer Aktivitäten stellen

(z.B.: Was genau wird mit den Subtests zum Hörverstehen überprüft und in welchem Verhältnis steht das Überprüfte zu den Skalen des GER?). Zur Illustration dieses deskriptiven Apparats seien hier einige wenige Beschreibungsaspekte herausgegriffen und auf das ZD j angewandt. In Bezug auf das Formular A1 des “Manuals” (“General Examination Description”) lässt sich für das ZD j etwa das Folgende festhalten:

- Was die “*Domänen*” (‘Lebens- und Sprachhandlungsbereiche’, vgl. GER, 52f.) betrifft, so spielen im ZD j besonders der private Bereich, Aspekte des Bildungswesens und z.T. auch der öffentliche Bereich eine wichtige Rolle, d.h. es wird davon ausgegangen, dass die Prüfungsteilnehmenden v.a. in diesen Domänen sprachlich handeln (müssen), und entsprechend werden die Situationen, Sprech- und Schreibanlässe, Themen und Texte der Prüfung im Hinblick auf diese Domänen ausgewählt und ‘arrangiert’.
- Die folgende Tabelle zeigt, welche *kommunikativen Aktivitäten* im ZD j vorkommen und gleichzeitig auch, wie sie überprüft und gewichtet werden. Während die Domänen im ZD und im ZD j anders gewichtet werden, gilt die folgende Darstellung unterschiedslos für beide Prüfungen:

Kommunikative Aktivitäten/ Prüfungsteile	Subtests	Aufgabentypen/ Antwortformate	Zeit	*relatives Gewicht
Leseverstehen	LV1: Globalverstehen	Zuordnung	90'	25%
	LV2: Detailverstehen	Mehrfachauswahl		
	LV3: Selektives Verstehen	Zuordnung		
Sprachbausteine	SB1: Lexik; Schwerpunkt Morphosyntax	Mehrfachauswahl		10%
	SB2: Lexik; Schwerpunkt Semantik/Pragmatik	Mehrfachauswahl		
Hörverstehen	HV1: Globalverstehen	Richtig-Falsch	ca. 30'	25%
	HV2: Detailverstehen	Richtig-Falsch		
	HV3: Selektives Verstehen	Richtig-Falsch		
Schriftliche Interaktion	Ein Subtest	(i.d.R. informeller) Brief als Antwort auf schriftliche Vorgabe (Brief o.ä.); 4 Leitpunkte bearbeiten	30'	15%

Mündliche Interaktion (mit Elementen mündlicher Produktion in den Subtests 1 u. 2)	Drei Subtests: 1: Kontaktaufnahme 2: Gespräch über ein Thema 3: Gemeinsam eine Aufgabe lösen	Zwei Varianten: <u>Paarprüfung</u> : Interaktion mit Partner/in <u>Einzelprüfung</u> : Interaktion mit Interlokutor	ca. 15'	25%
--	---	---	---------	-----

Abb. 2: Kommunikative Aktivitäten und Subtests im ZD j¹²

*Max. mögliche Punktzahl pro Subtest in Prozenten total erreichbarer Punkte.

- Die Bestehensgrenze ist beim ZD j gleich definiert wie beim ZD: Erreicht werden müssen 60% der möglichen Punkte, und zwar sowohl bei der schriftlichen Prüfung (Lese- und Hörverstehen, Sprachbausteine, schriftliche Interaktion) als auch bei der mündlichen Prüfung (mündliche Interaktion).

Erfolgreiche Prüfungen werden mit einer *Zertifikatsurkunde* bestätigt. Abgedruckt ist auf den Zeugnissen einerseits der Name und das GER-Niveau der Prüfung, andererseits werden die Prüfungsergebnisse aufgelistet, und zwar auch separat nach Prüfungsteilen. Das geht in Richtung eines Kompetenzprofils, das den LernerInnen auch eine gewisse Orientierung für das Weiterlernen ermöglicht.

- Zur Information und Unterstützung der Akteure sind für das ZD j, analog zum ZD, eine ganze Reihe von *Materialien* geplant, welche die neue Prüfung begleiten sollen, darunter: Informationstexte über die Prüfung, kommentierte Modellsätze (einschliesslich der Bewertungskriterien für schriftliche Leistungen und inklusive Musterkorrekturen) und kommentierte Videos, welche die Formate der mündlichen Aufgaben illustrieren und mündliche Leistungen sowie deren Bewertung veranschaulichen sollen.

Genauer abschätzen lässt sich das vermutete Niveau einer Prüfung mit den Formularen A9- A22 des "Manuals". Bei diesen Formularen geht es darum, die einzelnen Subtests der Prüfung und die damit überprüften kommunikativen Aktivitäten zu beschreiben und mit den Skalen des Referenzrahmens in Beziehung zu setzen. Spielt man das am Beispiel der ZD j Subtests zum

12 Zum Vergleich: Im Preliminary English Test (PET), die B1-Englisch-Prüfung von Cambridge, hat das Schreiben mehr Gewicht (25%), und die Grammatik tendenziell auch, obwohl sie sich nicht in dieser einfachen Weise quantifizieren lässt: Beim PET gibt es keinen eigenen Subtest für die Grammatik (ZD und ZD j: "Sprachbausteine"), aber die Aufgaben in den insgesamt 5 Leseteilen (25%) zielen oft auf Wörter und Strukturen. Ansonsten sind die Verhältnisse ähnlich, wobei der PET 'nur' 130 Minuten dauert, wohingegen für das ZD j und das ZD 165 Minuten veranschlagt sind.

Hörverstehen durch (Formular A9, "Listening Comprehension"), ergibt sich etwa das Folgende:

- Was weitere Spezifikationen der zielsprachlichen Verwendungskontexte betrifft (zusätzlich zu den *Domänen*, siehe oben, vgl. GER, 54), orientiert sich das ZD j primär an den sprachlichen Handlungsmustern bzw. "Szenarien", Strategien, Notionen und Themen, die für die Stufe des ZD erarbeitet wurden (LZB, S. 26 und 61ff.), wobei diese Lernzielkataloge für die Zielgruppen des ZD j teilweise adaptiert werden mussten oder noch müssen.
- Merkleich andere Akzente müssen im ZD j besonders auch bei den *Themen* gesetzt werden (vgl. auch GER, 58f.). Als Resultat aus umfangreichen Erprobungen zu diesem Aspekt ergab sich eine Art Themenhitparade mit den Spitzenreitern "Freizeit und Unterhaltung", "Medien" und "Ferien und Reisen", wohingegen Themen(felder) wie "Essen und Trinken", "Natur und Umwelt" und "Geschäfte und Konsum" relativ schlecht wegkamen (Genauerer zur Erprobung vgl. Kap. 4.4).
- Die *kommunikativen Aufgaben* und *Aktivitäten* des ZD j sind analog zum ZD gestaltet, wobei u.a. auf möglichst jugendgerechte Situierungen geachtet wird (vgl. auch GER, 59f., 71ff. und 153ff.). Das bedeutet z.B., dass die Jugendlichen ein Interview am Radio mit einem Rockstar hören und dabei einzelne Aussagen verstehen bzw. verifizieren oder falsifizieren müssen (HV2). Oder es bedeutet, dass sie die Hauptaussagen von eingespielten Statements aus einer Umfrage, z.B. zum Thema Noten, verstehen (HV1), oder spezifische Informationen in einer Durchsage im Schwimmbad heraushören müssen (HV3).
- Damit ist auch bereits angedeutet, welche *Textsorten* im Prüfungsteil Hörverstehen des ZD j hauptsächlich vorkommen (vgl. auch GER, 97): Im Vordergrund stehen Interviews, Statements und Kurzmeldungen am Radio sowie Anrufbeantwortertexte und Lautsprecherdurchsagen.

Teile dieser Spezifikationen lassen sich nun in den Skalen des GER zum Hörverstehen 'wiederfinden'. Wenn das der Fall ist, ergeben sich weitere Indizien für das Niveau eines Subtests und/oder Prüfungsteils. Für den Prüfungsteil Hörverstehen im ZD j sind in diesem Sinn besonders drei GER-Skalen ergiebig: "Hörverstehen allgemein", "Ankündigungen, Durchsagen und Anweisungen verstehen" und "Radiosendungen und Tonaufnahmen verstehen" (GER, 71ff.). In Bezug auf diese Skalen ergeben sich schon recht deutliche Hinweise darauf, dass die Hörverstehensaufgaben im ZD j im

Bereich von B1 liegen, und zwar, Insider wird es wenig überraschen, *im unteren Bereich von B1*. Beispielsweise lautet die Beschreibung für das Niveau B1.1 in der GER-Skala “Hörverstehen allgemein” (71f.): “Kann die Hauptpunkte verstehen, wenn in deutlich artikulierter Standardsprache über vertraute Dinge gesprochen wird, denen man normalerweise bei der Arbeit, in der Ausbildung oder der Freizeit begegnet; kann auch kurze Erzählungen verstehen.” Diese Beschreibung passt in der Tat ziemlich genau zum oben illustrierten Subtest HV2 des ZD j. Also liegt zumindest dieser Subtest auf dem Niveau B1.1? Man kann es nicht genau sagen, denn: Der Inputtext dieses Subtests wird zweimal eingespielt, und was das genau für das Niveau einer Aufgabe bedeutet, bleibt auch im “Manual” offen. Zudem wird die Schwierigkeit einer HV-Aufgabe natürlich durch eine Vielzahl von Faktoren beeinflusst, darunter text- und lernerseitige Faktoren ebenso wie solche der Aufgabenstellung (Instruktion; ‘Ablenker’ bei Auswahlaufgaben) und der Durchführung. Allerdings bleibt das “Manual” nicht bei Vorschlägen zur Analyse kommunikativer Aktivitäten stehen, sondern schlägt ein analoges Vorgehen zur genaueren Betrachtung von Aspekten der Sprachkompetenz vor, die zur Bewältigung kommunikativer Aufgaben nötig sind, bei Hörverstehensaufgaben z.B. Aspekte des ‘passiven’ Wortschatzes¹³. D.h. nichts anderes, als dass die Analyse weitergehen und schon für das Hörverstehen ‘verdoppelt’, und dann natürlich für weitere Prüfungsteile in gleicher Weise fortgesetzt werden müsste. Entsprechend lautet das (vorläufige) Fazit dieser ‘Übung’: Mit Hilfe des Spezifikations-Teils des “Manuals” lassen sich in der Tat genauere und v.a. zeigbare Anhaltspunkte für die Niveauezugehörigkeit einer Prüfung erarbeiten; das ist nicht wenig, und doch sind es nur bessere Anhaltspunkte¹⁴.

13 Die entsprechenden Kategorien des GER sind: Linguistische Kompetenzen, Soziolinguistische Kompetenzen, Pragmatische Kompetenzen und, soweit nicht schon in der Analyse der rezeptiven Aktivitäten berücksichtigt, Strategische Kompetenzen (ebd., 109ff. und 71f.).

14 Ganz ausser Acht gelassen wurde in diesen Betrachtungen, dass die kalibrierten Deskriptoren, aus denen die Niveaubeschreibungen des GER bestehen, aus empirischen Untersuchungen kommen, die vorwiegend mit *Erwachsenen* durchgeführt wurden (vgl. Schneider & North 2000). Fragen der Adaptation von Deskriptoren für die Zielgruppe Kinder und Jugendliche spielen im Rahmen des Deutschschweizer Schulprojekts IEF (Entwicklung von Instrumenten für die Evaluation von Fremdsprachenkompetenzen) eine grosse Rolle. IEF wird am LeFoZeF der Universität Freiburg von Peter Lenz und Thomas Studer bearbeitet; Projektkoordination: Bildungsplanung Zentralschweiz.

4.3. Zur Begründung des ZD j

Es sind im Wesentlichen zwei Gründe, die zur Entwicklung des ZD j geführt haben; der eine betrifft die Seite des Prüfungsangebots, der andere ist an Bedürfnissen von Lernenden und Schulen orientiert.

In Bezug auf das *Angebot* lässt sich die Situation für DaF so darstellen:

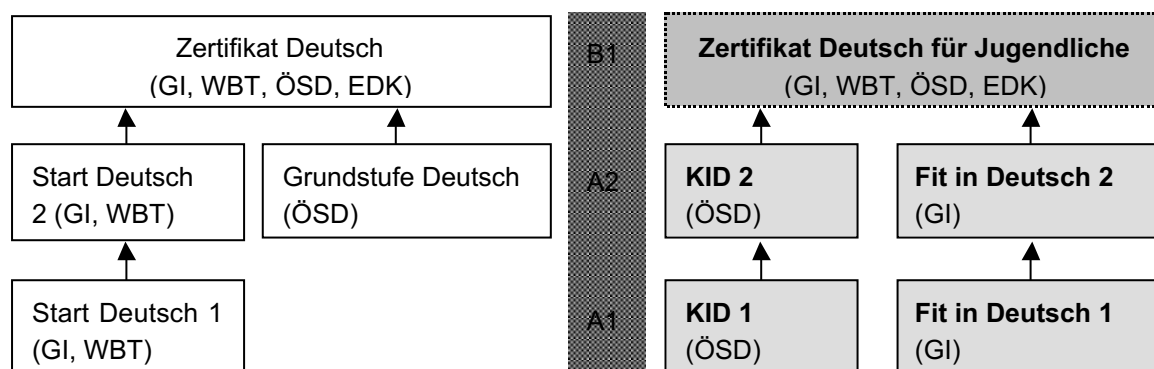


Abb. 3: Zertifikate für DaF im Bereich der Niveaus A1 bis B1

Links: Zertifikate für Erwachsene; rechts: Zertifikate für Kinder und Jugendliche

Abb. 3 zeigt, dass das ZD j eine 'systematische' Lücke auf der Landkarte internationaler DaF-Zertifikate schliesst: Bis anhin kam für jugendliche LernerInnen, die bereits über eine Sprachkompetenz im Bereich von B1 verfügen und die ihre Kompetenzen mit einem internationalen Diplom nachweisen wollten, nur das ZD Frage. Damit mussten sie eine Prüfung ablegen, die für Erwachsene konzipiert ist, d.h. sie wurden mit Situierungen, Inhalten, Sprechweisen und Instruktionen konfrontiert, die zumindest in Teilen ihrem Erfahrungshorizont und Lernkontext nicht angemessen sind.

Von Interesse ist in diesem Zusammenhang ein Vergleich mit Englisch: Im Rahmen der Cambridge-Diplome gibt es das YLE-Programm (Young Learners English), das die drei Prüfungen "Starters" (für ca. 7-Jährige, Stufe unter A1), "Movers" (ca. 8-11-Jährige, A1) und "Flyers" (ca. 9-12-Jährige, A2) umfasst. Das Flyers-Diplom ist der ersten Erwachsenenprüfung, dem KET (Key English Test), hinsichtlich des attestierten Niveaus äquivalent; beide Diplome sind auf dem Niveau A2 situiert¹⁵. Demgegenüber geht Deutsch einen etwas andern Weg: Ein Pendant zu den "Starters" gibt es (noch?) nicht, wohl aber kinder- und jugendorientierte Prüfungen für A1/A2 (Fit in Deutsch 1/2, KID 1/2), die hinsichtlich des Niveaus und der ungefähren Altersangaben mit den "Movers"

¹⁵ Vgl. http://www.cambridge-efl.org/support/downloads/yle/yle_hb_03.pdf: YLE-Handbuch, Ausgabe 2003, S. 4.

bzw. mit den “Flyers” vergleichbar sind. Und zusätzlich liegt jetzt dann mit dem ZD j eine B1-Prüfung für Jugendliche zwischen 12 und 16 Jahren vor, die kein Gegenstück im YLE-Programm hat.

Was die Seite der Nachfrage betrifft, so gibt es offenbar ein *Bedürfnis* für jugendorientierte DaF-Zertifikate auf dem Niveau B1, im Ausland z.B. in Griechenland, wo Fremdsprachenzertifikate traditionell eine grosse Rolle spielen, aber auch im deutschsprachigen Inland (speziell zur Schweiz siehe unten, 4.5), etwa im Rahmen von Sommersprachkursen oder z.B. auch in nicht-deutschsprachigen Privatschulen. Darüber hinaus scheint im Inland ein zunehmendes Interesse daran zu bestehen, Fremdsprachenzertifikate in Schulabschlüsse an öffentlichen Schulen zu integrieren. Zwei Beispiele aus Deutschland: 2001 haben in Baden-Württemberg 500 Realschüler die Cambridge-B1-Prüfung PET abgelegt (Weskamp 2003, 140). Die WBT kooperiert seit einigen Jahren mit dem Saarland: Dort ist es den SchülerInnen seit 2003 möglich, im Rahmen der zentralen Prüfung für den mittleren Bildungsabschluss (10. Schuljahr) ein TELC-B1-Zertifikat in Französisch und Englisch zu erwerben.

4.4. Zum Projektstand

In der Entwicklung und Erprobung des ZD j lassen sich grob zwei Phasen unterscheiden:

In einer *ersten Phase* wurden, ausgehend v.a. von den Lernzielbestimmungen des ZD (EDK et al., Hrsg., 1999), Fragenkataloge mit den zu adaptierenden Punkten erstellt und diskutiert.

Neben den oben (4.2) erwähnten Anpassungen war u.a. auch ‘die’ Jugendsprache ein Thema, wobei rasch klar war, dass sich dieses Thema aus Sicht der Testkonstruktion im Wesentlichen auf die Frage nach den Inputtexten einschränken lässt, und selbst in diesem Bereich schien Augenmass angezeigt, denn auch Jugendliche (Muttersprachler ebenso wie Fremdsprachige) sind ja oft mit Texten konfrontiert, die zumindest auch an Erwachsene gerichtet sind. Eine Konsequenz daraus ist, dass gruppen- oder jugendsprachliche Varianten in den Hör- und Lesetexten sowie in den Sprachbausteinen des ZD j schon dann und wann vorkommen (z.B. in narrativen Texten und dort v.a. auf der Ebene der Lexik), aber sehr moderat. Wegweisend und für die Auswahl von Texten sehr hilfreich waren in diesem Zusammenhang besonders auch Gespräche, die mit der Jugendsprachforscherin Eva Neuland geführt werden konnten.

Parallel zu diesen Adaptationsüberlegungen haben alle vier Kooperationspartner je einen Entwurf für einen Testsatz erarbeitet und in einen Pool eingebracht, aus dem ein vorläufiger Modellsatz und zwei provisorische Prüfungssätze erstellt wurden. In einem nächsten Arbeitsschritt wurden die vorläufigen Testsätze auch mit LehrerInnen, die über viel Erfahrung mit der Zielgruppe verfügen, besprochen und unter Beizug der für das ZD erarbeiteten Konstruktionsrichtlinien für die Erprobung vorbereitet.

Bei der Erprobung des Modelltests wurden zwei Akzente gesetzt: Zum einen sollte der Schwierigkeitsgrad der Prüfung umfassender abgeklärt werden, weshalb neben dem Modelltest auch ein sog. Ankertest¹⁶ zum Einsatz kam. Zum andern ging es darum, mehr über die verschiedenen Adaptationen und ihre Wirkung auf Prüfungsteilnehmende und Lehrpersonen zu erfahren. Zu diesem Zweck wurde ein Fragebogen erstellt und ebenfalls parallel zum Modelltest in Umlauf gebracht. Der Fragebogen war eng auf den Modelltest bezogen, fokussierte u.a. die Aspekte Instruktionsverständnis, Inhalt und Sprache der Prüfungsteile und Subtests und zirkulierte in drei Versionen: Eine ausführliche Version richtete sich an MuttersprachlerInnen, eine reduzierte an Fremdsprachige und eine nochmals etwas andere an Lehrpersonen.

An der Erprobung nahmen insgesamt 416 Jugendliche (236 Fremdsprachige und 180 MuttersprachlerInnen) und mehr als zwei Dutzend Lehrpersonen teil, wobei nicht bei allen ProbandInnen alle Erprobungsinstrumente eingesetzt werden konnten. Die Stichprobe der Fremdsprachigen war breit gestreut und reichte von Wien über Paris, Barcelona und Győr (Ungarn) bis nach Varna (Bulgarien) und Blumenau (Brasilien); die MuttersprachlerInnen kamen aus verschiedenen Regionen der Deutschschweiz.

Auf diesem Weg wurden im Herbst 2003 viele Daten erhoben, die mit qualitativen (z.B. Durchsicht von Briefen mit Bezug auf GER-Skalen) und besonders auch mit verschiedenen quantitativen Methoden ausgewertet wurden. Die Auswertung der 'harten' Daten aus dem Modelltest erfolgte einerseits mit dem weit verbreiteten Item- und Testanalyse-Programm "Iteman" und andererseits mit einem Programm für die Rasch-Analyse ("Raquel for Windows"), mit dem auch die Resultate des Ankertests analysiert

16 Der Ankertest bestand aus isolierten Wortschatz- und Grammatik-Items (Mehrfachauswahl-Aufgaben) mit bekanntem und dem Niveau B1 des GER zugeordnetem Schwierigkeitsgrad. Entwickelt, durchgeführt und ausgewertet wurde der Ankertest ausserhalb der trinationalen Kooperation unter Beteiligung des GI.

wurden; die Fragebogendaten wurden mit allgemeineren statistischen Methoden ausgewertet, darunter Varianzanalyse und Rangkorrelation.

Die statistischen Befunde der Iteman-Analysen erlaubten es, problematische Items (z.B. solche mit zu geringer Trennschärfe) zu identifizieren. Die Ergebnisse der Rasch-Analysen wiesen darauf hin, dass der Modelltest zu leicht sein könnte. Die Fragebogendaten wurden v.a. herangezogen, um mögliche Gründe für nicht oder schlecht funktionierende Aufgaben zu erkennen (in einem Fall erwies sich z.B. eine Instruktion als problematisch). Zwar wiesen einige der Befunde auch in verschiedene Richtungen, z.B.: War vielleicht nicht der Test zu leicht, sondern könnte auch, wofür die Briefanalyse Anhaltspunkte bot, ein Teil der ProbandInnen "zu gut" gewesen sein? Insgesamt aber sprachen die Erprobungsergebnisse schon eine klare Sprache: Positive Befunde auf der einen Seite (z.B. konnte gezeigt werden, dass das Verständnis der Instruktionen zu den Prüfungsaufgaben statistisch unabhängig ist von der Prüfungserfahrung¹⁷ und auch vom Alter der ProbandInnen), und beträchtlicher Handlungsbedarf auf der andern (z.B. gab es nicht wenige Items mit mangelhafter Trennschärfe).

Aufgrund dieser Befunde wurde der gesamte Modelltest einer umfassenden Revision unterzogen; ein Subtest (die Briefaufgabe) wurde sogar ganz ausgetauscht. Weil alle Subtests modifiziert und z.T. erheblich verändert wurden, wurde eine zweite Erprobung nötig. Bei dieser Erprobung (mit nur noch ca. 150 fremdsprachigen ProbandInnen) läuft wiederum ein Ankertest mit, nicht aber der Fragebogen, der seine Zwecke erfüllt hat. Die Ergebnisse dieser *zweiten Phase der Erprobung* werden demnächst vorliegen und in den definitiven Modelltest des ZD j eingearbeitet werden können.

4.5. Überlegungen zum Einsatz des ZD j in der Schweiz

Mit dem ZD j liegt demnächst eine standardisierte Sprachprüfung für Jugendliche im Alter von 12 bis 16 Jahren vor, die auf dem Niveau B1 situiert ist. Wenn man davon ausgeht, dass Schweizer SchülerInnen am Ende der obligatorischen Schulzeit in den Fremdsprachen dieses Niveau erreichen

17 Beispiel einer einfaktoriellen Varianzanalyse aus den Fragebogendaten für den Zusammenhang zwischen Prüfungserfahrung (Alternativfrage) und Instruktionsverständnis (dreistufiges semantisches Differential) beim Subtest LV1: n=68; DF=1, Mean Diff.=0,005, Scheffe F-Test=0,002, p=0,965.

können¹⁸, dann steht die Frage im Raum, ob das ZD j auch an öffentlichen Schulen angeboten werden sollte, z.B. als Teil von Abschlussprüfungen auf der Sekundarstufe I in den französisch- und italienischsprachigen Landesteilen und etwa auch als Option für Aufnahmeprüfungen an den dortigen Gymnasien.

Die hier vertretene These ist, dass eine standardisierte Sprachprüfung wie das ZD j durchaus eine sinnvolle, für alle beteiligten Akteure nützliche und wertvolle Ergänzung zum bestehenden Prüfungswesen an öffentlichen Schulen sein kann. Für den Einbezug standardisierter Prüfungen spricht v.a. auch, dass sie, wie Glaboniat (2002) überzeugend zeigen kann, schulischen Tests überlegen sind, was die Berichts- und die Selektionsfunktion von Leistungsbeurteilungen betrifft. Die Begründungen für diese Überlegenheit ergeben sich zum grösseren Teil bereits aus dem Abschnitt über Charakteristika von Zertifikaten (vgl. oben, 1.2, und besonders Glaboniat 2002, 224f):

- *Berichtsfunktion*: Schulische Tests sind in der Regel normorientiert (Gruppennorm) und haben deswegen zwar einen gewissen Informationswert für direkt involvierte SchülerInnen, Lehrpersonen und Eltern, welche die Gruppe und den Lernkontext genau kennen, aber “nach aussen” ist dieser Informationswert äusserst gering. Demgegenüber haben kriterienreferenzierte Sprachprüfungen wie das ZD j durch ihre transparenten Bezüge auf genau beschriebene Kompetenzen eine hohe Aussagekraft (auch) für Aussenstehende.
- *Selektionsfunktion*: Selektionsprozesse, die auf schulische Tests abgestützt sind, sind sehr problematisch, weil die Gültigkeit dieser Tests in sehr vielen Fällen stark eingeschränkt ist. Schulische Tests berücksichtigen häufig nur leicht erfassbare Komponenten von Sprache wie Wortschatz und Grammatik (z.B. Zählen formalsprachlicher Fehler in Schreibaufgaben). Ein Gesamturteil über Sprachkompetenz, das darauf

18 Welches Niveau Schweizer SchulabgängerInnen in den Fremdsprachen genau erreichen können und sollen, wird im Rahmen des Projekts HarmoS noch genauer zu untersuchen und festzulegen sein (vgl. dazu Maradan in diesem Heft). Deutliche Signale dafür, dass das Niveau B1 (vorab in einer zweiten Landessprache) politischer Konsens sein kann, gibt es indessen schon seit längerem (vgl. das Gesamtsprachenkonzept und verschiedene andere Pressemitteilungen der EDK; Literaturhinweise im Beitrag von Maradan). Und es gibt auch erste empirische Indizien dafür, dass dieses Niveau erreichbar ist: In einer (unveröffentlichten) Pilotstudie im Rahmen von IEF (vgl. Fussnote 14) wurden in zwei 3. Sekundarschulklassen (9. Schuljahr) Teile des *Certificat Français*, der TELC-B1-Prüfung für Französisch, eingesetzt. Die Ergebnisse lassen vermuten, dass 28 von 30 beteiligten SchülerInnen gute Aussichten gehabt hätten, diese Prüfung zu bestehen.

(und vielleicht noch auf andere Faktoren wie z.B. "Mitmachen" im mündlichen Unterricht abgestützt ist), ist aus testtheoretischer Sicht nicht valide, weil nicht klar ist, was genau das mit Sprachkompetenz zu tun hat. Demgegenüber versuchen kommunikative Prüfungen wie das ZD j die Validität der Aufgaben dadurch sicherzustellen, dass die Aufgaben in struktureller Analogie zu realen Sprachgebrauchssituationen (bzw. zu einem Modell davon) konstruiert werden (vgl. dazu das Authentizitätskonzept von Bachman & Palmer (1996 10ff., 23).

Problematisch ist bei schulischen Tests in der Regel auch die Zuverlässigkeit der Auswertung ("scorer reliability"). Gerade bei offeneren Aufgabenformen wie z.B. dem Schreiben von längeren Texten zeigt sich immer wieder, dass verschiedene Beurteilende zu sehr unterschiedlichen Noten kommen, weil nicht klar ist, wie die Rohwerte ermittelt werden sollen (Was ist ein Fehler?) und ebensowenig, wie diese Werte in Punkte und Noten umzurechnen sind (Was ist das relative Gewicht eines Fehlers?). Um möglichst reliable Bewertungen zu erreichen, werden bei standardisierten Prüfungen eine ganze Reihe von Massnahmen getroffen. Bei Schreibaufgaben reichen diese von präzisen Aufgabenstellungen und genauen Arbeitsanweisungen über detaillierte und skalierte Bewertungskriterien bis hin zur Schulung von BewerterInnen anhand von Musterkorrekturen.

Würde man die Berichts- und die Steuerungsfunktion standardisierten Prüfungen überantworten, käme das nicht zuletzt einer bedeutenden Entlastung schulischer Tests gleich, denn diese müssen ja v.a. auch *pädagogische Funktionen* erfüllen, und da haben sie durchaus Vorteile gegenüber Zertifikaten: Dadurch, dass sich Zertifikate gerade nicht auf spezifische Lehr-Lern-Kontexte beziehen, ist ihr Nutzen als unmittelbares Feedback- und Steuerungsinstrument für den Unterricht eher beschränkt. Hier, bei der Überprüfung von spezifischen Lernfortschritten und zur Optimierung von Unterrichtssequenzen können schulische Tests durchaus eine wichtige Rolle spielen.

Zusammenfassend lässt sich daher sagen: Standardisierte Prüfungen können und sollen schulische Tests nicht ersetzen, aber Zertifikate wie das ZD j könnten schulische Tests auf sehr sinnvolle Weise ergänzen.

Ein weiteres Argument für die Integration von standardisierten Prüfungen in das schulische Prüfungswesen ist der oben (2.2) diskutierte Washback-Effekt: Obwohl dieser Effekt empirisch nur schwer fassbar ist, kann man davon ausgehen, dass zielgruppenspezifische kommunikative Prüfungen eine posi-

tive Rückwirkung auf das Lernen und den Unterricht haben, indem z.B. im Unterricht vermehrt solche Aufgaben zum Zug kommen, die an die Formate der Prüfung angelehnt sind (vgl. auch Weskamp 2003, 138¹⁹).

Auf der andern Seite besteht eine gewisse Gefahr, dass beim Einsatz standardisierter Prüfungen andere schulische Lernziele, darunter spielerischer und kreativer Sprachgebrauch oder das sog. Fremdverstehen, die mit diesen Prüfungen nicht erfasst werden, in den Hintergrund geraten. Hier käme es v.a. auch auf eine Fein-tuning der verschiedenen Lernziele an.

Besondere Vorsicht ist ferner v.a. auch im Zusammenhang mit der Selektionsfunktion von Prüfungen geboten: Selektionsentscheide sollten nicht nur auf ein Zertifikat, sondern durch eine umfassende Leistungsbeurteilung abgestützt sein, und dazu eignet sich besonders auch das Europäische Sprachenportfolio (ESP). Was die Dokumentationsfunktion des ESP betrifft, so haben Zertifikate dort einen wichtigen Platz, und gleichwohl sind sie nur ein Element unter anderen: Selbstbeurteilung, die ja auf die gleichen Kompetenzskalen Bezug nimmt wie die Zertifikate, und ausgewählte Arbeiten, die typische Lernerleistungen illustrieren, sind dort ebenso wichtig. Insofern können sich nicht nur Zertifikate und schulische Tests, sondern v.a. auch standardisierte, auf die Kompetenzskalen des GER bezogene Prüfungen wie das ZD j und das ESP auf sehr sinnvoller Art und Weise ergänzen.

In der Schweiz sind die Vorzeichen für derartige Synergien günstig, denn demnächst geht das ESP II, das sich an 11-15-Jährige richtet, in Erprobung.

LITERATUR

Alderson, J.C. & Banerjee, J. (2001/2002). Language testing and assessment. Part I: *Language Testing* 34, 213-236. Part II: *Language Testing* 35, 79-113.

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

Bachman, L.F. & Cohen, A.D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

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

19 Weskamp (ebd.) betont allerdings mit Recht, dass es für positiven Washback nicht ausreicht, wenn Lehrpersonen nur die Form von Klassenarbeiten anpassen und ihre Lehrmethodik beibehalten. Positiver Washback könne nur erreicht werden, wenn eine neue Überprüfungs-kultur Hand in Hand mit einer veränderten Lehr-Lernphilosophie gehe, und um das zu erreichen, müsse den Lehrpersonen v.a. auch das Konstrukt der neuen Prüfung deutlich gemacht werden.

- Baldegger, M., Müller, M. & Schneider, G. (1980). *Kontaktschwelle Deutsch als Fremdsprache*. Strassburg: Langenscheidt.
- Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Berlin et al.: Springer.
- Bygate, M., Skehan, P. & Swain, M. (2001). *Researching pedagogic tasks. Second language learning, teaching and testing*. Harlow: Pearson Education.
- Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual*. Preliminary Pilot Version, November 2003. Strasbourg.
- Davies, A. (2003). Three heresies of language testing research. *Language Testing* 20, 355-368.
- Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Hrsg. vom Goethe-Institut, der KMK, der EDK und dem BMBWK. Berlin et al.: Langenscheidt. [zitiert als GER]
- Glaboniat, M. (1998). *Kommunikatives Testen im Bereich Deutsch als Fremdsprache*. Innsbruck et al.: Studienverlag.
- (2002). Schulnoten versus standardisierte Prüfungen – Gedanken zum Neben- und Gegeneinander schulischer und standardisierter Leistungsmessung im DaF-Bereich. In H. Barkowski & R. Faistauer (Hrsg.) *...in Sachen Deutsch als Fremdsprache* (S. 217-230). Hohengehren: Schneider Verlag.
- Goethe-Institut, ÖSD, EDK, WBT (Hrsg.) (1999): *Zertifikat Deutsch. Lernziele und Testformat*. Frankfurt a.M.: WBT. [zitiert als LZB]
- Gogolin, I. (2001). Sprachstandsdiagnosen. In G. Helbig, L. Götze, G. Henrici & H.-J. Krumm (Hrsg.) *Deutsch als Fremdsprache: ein internationales Handbuch* (2 Bde), Bd. 2 (S. 1007-1016). Berlin: de Gruyter.
- Krause, W.-D. & Sändig, U. (2002). *Testen und Bewerten kommunikativer Leistungen im Unterricht Deutsch als Fremdsprache*. Frankfurt a.M. et al.: Lang.
- Krumm, H.-J. (2003). Sprachenpolitik und Mehrsprachigkeit. In B. Hufeisen & G. Neuner (Hrsg.) *Mehrsprachigkeitskonzept – Tertiärsprachen – Deutsch nach Englisch* (S. 35-49). Strassbourg: Council of Europe Publishing.
- Kunnan, A. J. (Ed., 2000). *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.
- McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.
- Morfeld, P. (2003). Sprachenzertifikate. In K.-R. Bausch, H. Christ & H.-J. Krumm (Hrsg.) *Handbuch Fremdsprachenunterricht* (S. 384-387). Tübingen et al.: A. Francke Verlag.
- Schifko, M. (2001). Prüfungen, Zertifikate, Abschlüsse als Planungskategorien für den Unterricht. In G. Helbig, L. Götze, G. Henrici & H.-J. Krumm (Hrsg.) *Deutsch als Fremdsprache. Ein internationales Handbuch* (2 Bde), Bd. 1 (S. 827-834). Berlin: de Gruyter.
- Schneider, G. (2000). Lernerfahrungen reflektieren und Lernerfolge beurteilen. *Babylonia*, 1, 6-9.
- Schneider, G. & North, B. (2000). *Fremdsprachen können – was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Chur: Verlag Rüegger.
- Shohamy (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education.
- Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate et al. (Eds.) *Researching pedagogic tasks. Second language learning, teaching and testing* (S. 167-185). Harlow: Pearson Education.
- Studer, T. (2000). Internationale Sprachenzertifikate aus der Sicht von Schweizer TestautorInnen. *Babylonia*, 1, 62-66.

- (2001). Welches Deutsch soll geprüft werden? Fragen zu einem Sprachtest für Einbürgerungswillige. *NZZ*, 31, 45.
- Vollmer, H.J. (2003). Leistungsmessung, Lernerfolgskontrolle und Selbstkontrolle. In K.-R. Bausch, H. Christ & H.-J. Krumm (Hrsg.) *Handbuch Fremdsprachenunterricht* (S. 365-370). Tübingen et al.: A. Francke Verlag.
- Voss, B. (2001). Die UNlcert-Initiative: Ein Beitrag zu Transparenz und Qualitätssicherung im universitären Fremdsprachenbereich. In H. Pürschel & U. Raatz (Hrsg.) *Tests and translation* (S. 169-187). Bochum: AKS-Verlag.
- Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System* 28, 499-509.
- Weskamp, R. (2003). *Fremdsprachenunterricht entwickeln*. Hannover: Schroedel Verlag.

