Publié dans
Bulletin VALS-ASLA (Vereinigung für angewandte Linguistik in der Schweiz - Association suisse de linguistique appliquée)
89, 165-168, 2009
qui doit être utilisé pour toute référence à ce travail

Sprachkompetenzen messen in Schule und Wissenschaft – Reflexionen anlässlich der Lancierung eines Forschungsprojekts

Anton NÄF
Université de Neuchâtel, Institut de langue et littérature allemandes,
Espace Louis-Agassiz 1, CH-2000 Neuchâtel
anton.naf@unine.ch

Le présent article, consacré à la mesure de compétences langagières, met en évidence les similitudes et les différences entre tests informels (utilisés en classe) et tests standardisés (utilisés entre autres pour la recherche). Afin de concrétiser les réflexions d'ordre méthodologique, il présente les premiers résultats d'une étude longitudinale sur la maturité bilingue en Suisse. Cette recherche montre que les gymnasiens inscrits à cette filière obtiennent, dès le départ, des résultats de tests significativement supérieurs à ceux des élèves des classes parallèles servant de groupe témoin. L'article se termine par des considérations sur le débat actuel en rapport avec une éventuelle introduction de standards de formation dans le domaine des langues étrangères au niveau gymnasial.

Mots clés:
Maturité bilingue, immersion, tests de langue informels et standardisés

1. Einleitung

Die Durchführung von Lernerfolgskontrollen gehört zum Kern- und Alltagsgeschäft jeder Lehrperson, somit auch jedes Fremdsprachenlehrers. Die von den Schülern bei den entsprechenden Tests erzielten Leistungen haben unter anderem den Zweck, der Lehrperson die Vergabe einer möglichst objektiven und gerechten Note zu ermöglichen. Aber auch in der Sprachwissenschaft werden – nach dem Vorbild der empirischen Sozialwissenschaften – je länger je mehr Testverfahren als Messinstrumente eingesetzt. Die in der wissenschaftlichen Forschung zum Einsatz gelangenden Tests zur Messung der Sprachkompetenz unterscheiden sich jedoch stark von den in der Schule verwendeten.

Im Folgenden sollen nun im zweiten Abschnitt einige terminologische Klärungen vorgenommen werden, sodann soll im dritten der Frage nachgegangen werden, inwiefern und warum sich die Tests in Schule und Wissenschaft unterscheiden. Anschliessend werden im vierten Abschnitt die in der Neuenburger Longitudinalstudie zur zweisprachigen Maturität (2005-2008) zum Einsatz gelangten Testverfahren vorgestellt. Im fünften Abschnitt werden die bei der Eingangsmessung erzielten Resultate vorgestellt und interpretiert. Zum Schluss soll die im Titel angesprochene Problematik in den gegenwärtig laufenden bildungspolitischen Diskurs eingebettet werden.

2. Vorbemerkung zur Terminologie

Für das Phänomen, das hier zur Debatte steht, gibt es in der Fachliteratur eine grössere Zahl von – mehr oder weniger synonym verwendeten – Termini. In der vorliegenden Studie verwenden wir vor allem das Kompositum *Sprachstandsmessung*. Sowohl für dessen Vorderglied *Sprachstand* als auch für das Hinterglied *Messung* existieren konkurrierende Bezeichnungen, die jedoch ein zum Teil etwas anderes Licht auf das damit gemeinte Phänomen werfen.

Statt von Sprach*stand* wird auch von Sprach*niveau*, Sprach*kompetenz* oder Sprach*leistung* gesprochen. Wenn im Rahmen unserer Untersuchungen zur zweisprachigen Maturität dem Terminus Sprach*stand* der Vorzug gegeben wird, dann deshalb, weil er an die Frage "Wo stehe ich mit meinen Sprachkenntnissen?" anklingt und damit die Idee einer Bilanz transportiert, die zu einem bestimmten Zeitpunkt des Lernprozesses gezogen wird, und dies unabhängig von der Art und Weise des Spracherwerbs (Familie, Schule, Sprachaufenthalt usw.). Ebenfalls geeignet ist der Ausdruck Sprach*niveau*; zumindest tendenziell setzt er aber die Existenz einer entsprechenden Messskala voraus (etwa derjenigen des Gemeinsamen Europäischen Referenzrahmens GER). Von Vorteil ist, dass er ein Pendant in andern Sprachen hat, z.B. *niveau de langue* im Französischen. Der im Englischen dominierende Terminus scheint dagegen immer noch *language proficiency* zu sein, dies trotz aller Kritik, welcher die diesem Begriff zugrundeliegenden theoretischen Grundannahmen in den letzten Jahrzehnten ausgesetzt waren[1].

Einer etwas ausführlicheren Begründung bedarf die Wahl des Hinterglieds *Messung*. Messen bedeutet, 'etwas in seinen Massen, in seiner Grösse bestimmen'. Bei seinem prototypischen Gebrauch geht es dabei stets um quantitativ erfassbare Grössen, z.B. ein Brett (mit dem Metermass) messen, die Zeit (mit der Stoppuhr) messen, das Fieber (mit dem Thermometer) messen oder den Luftdruck (mit dem Barometer)[2] messen. Seit längerem wird auch in den Sozialwissenschaften von Messen gesprochen. So wird etwa mit bestimmten Tests die Intelligenz gemessen, genauer gesagt der sogenannte Intelligenzquotient. Man muss sich dabei aber stets vor Augen halten, dass es sich hier in aller Regel nicht um Intervallskalen handelt. Wie anderswo handelt es sich auch in der Spracherwerbsforschung meist nicht um Intervalldaten, sondern um Messen im Sinne des Vergleichs von Daten – etwa der erreichten Punktzahl in einem Test – auf einer Ordinalskala ("x ist grösser als y"). Mit

---

1 Zu den Versuchen, die beiden Konzepte *proficiency* und *communicative competence* theoretisch und für die Testpraxis gegeneinander abzugrenzen, vgl. McNamara, 1996: 76-79.

2 Auf diese Metaphorik rekurriert der für die Bestimmung des Grads der Mehrsprachigkeit der Europäer verwendete Ausdruck Eurobarometer (EUROBAROMETER 54 Special "Europeans and Languages" 2001).

andern Worten: Messen wird hier in einem metaphorischen Sinn verwendet. Streng genommen müssten wir also im Folgenden "messen" und "Messung" immer zwischen Anführungszeichen setzen. Wir wollen aber darauf verzichten, weil sich der Ausdruck nicht bloss im Deutschen, sondern auch in andern Sprachen (engl. *measurement*, franz. *mesure*) in den Sozialwissenschaften etabliert hat. Konkurrierende Termini sind insbesondere die beiden Internationalismen *Diagnose* und *Evaluation*. Diagnose ist aber noch stark seinem medizinischen Herkunftsbereich verhaftet, und Evaluation suggeriert über eine blosse punktuelle Messung hinaus eine umfassende Bestandesaufnahme und Bewertung der Ergebnisse.

Neben dem hier favorisierten Terminus *Sprachstandsmessung* (engl. *language proficiency measurement*, franz. *mesure du niveau de langue*) werden – in Form eines kompakten Kompositums oder einer Nominalgruppe – zur stilistischen Variation gelegentlich auch Bezeichnungen wie Messung des Sprachniveaus, Erhebung des Sprachstands u. Ä. verwendet.

3. Informelle vs. standardisierte Tests

Auch wenn es nicht möglich sein dürfte, zwischen den in der Schule praktizierten informellen Tests einerseits und den bei den internationalen Sprachenzertifikaten (*University of Cambridge, Alliance française, Goethe-Institut*) und in der wissenschaftlichen Forschung zum Einsatz gelangenden standardisierten Tests andererseits eine absolut trennscharfe Grenze zu ziehen, so sind die beiden Typen von Tests in der Regel doch hinreichend verschieden voneinander, nicht nur bezüglich der Art der gestellten Aufgaben, sondern auch – und vielleicht vor allem – hinsichtlich der Rahmenbedingungen ihrer Durchführung.

Eine terminologische Vorbemerkung: *Test* und *Prüfung* werden hier synonym verwendet. In der Fachliteratur wird der Ausdruck Prüfung zum Teil nur für im Rahmen des regulären Schulbetriebs durchgeführte Leistungskontrollen verwendet, und zumindest tendenziell meint Prüfung etwas Umfassenderes als Test (Perlmann-Balme, 2001: 996).

Auch wenn wir es bei der Unterscheidung informell vs. standardisiert letztlich offenbar eher mit einem Kontinuum als mit einer strikten Dichotomie zu tun haben, ist eine polare Unterscheidung der beiden Konzepte – wie sie in der Tabelle 1 mit Hilfe von zwölf Gegensatzpaaren vorgeschlagen wird – durchaus gerechtfertigt. Weniger im Sinne der klassischen Definitionspraxis als vielmehr in jenem der Prototypen-Semantik (es gibt jeweils "beste Vertreter" für jede der beiden Kategorien) macht eine solche Trennung unseres Erachtens durchaus Sinn.

INFORMELLE TESTS	STANDARDISIERTE TESTS
mit Unterrichtsbezug (zuvor Behandeltes)	ohne Unterrichtsbezug
häufig, gehört in der Schule zum Alltag; auf wenige Phänomene konzentriert	selten, besondere Gelegenheiten; umfassender, umfangreicher, zeitaufwendiger
Lernfortschrittstest *(progress test)*; formative Evaluation; begrenzte, kleinschrittige Lernziele	Leistungstest *(performance test)*[3]; summative Evaluation; globaler Sprachbeherrschungsgrad
individuelle Leistung auf Gruppe von Lernern bezogen (z.B. Klassendurchschnitt); normbezogen	individuelle Leistung auf standardisierte Kompetenzkriterien bezogen (z.B. Lernziele); kriteriumsbezogen[4]
vom Lehrer für die betreffende Klasse gemacht *(teacher-made test, classroom-test)*	von spezialisierter Einrichtung (z.B. Test-Institut, Universität) entwickelt
meist auf ein bestimmtes Lehrmittel bezogen	kein Bezug zu einem bestimmten Lehrmittel
ohne vorgängige Erprobung; ad hoc zusammengestellt	an grosser Stichprobe geeicht; erfüllen in hohem Masse Gütekriterien wie Validität, Objektivität, Reliabilität
tendenziell "natürliche" Aufgaben	tendenziell "künstliche" Aufgaben
intern (Klasse, Schule); im normalen Klassenraum abgelegt	extern; in "fremden" Räumen abgelegt
obligatorisch für alle Schüler einer Klasse	Teilnahme freiwillig
tendenziell rückwärts gerichtet	tendenziell zukunftsorientiert[5] (Zulassung, Anstellung, Beförderung, usw.)
keine Gebühren	meist kostenpflichtig

Tab. 1: Informelle vs. standardisierte Tests (polare prototypische Eigenschaften)

Ein idealtypisches Robotbild der beiden Typen von Tests sieht somit etwa folgendermassen aus: Ein *informeller Test*, z.B. eine Sprachprüfung im Fremdsprachenunterricht, ist in den regulären Unterrichtsverlauf eingebaut und bezieht sich auf den vorangegangenen Unterricht in einer bestimmten Klasse. Er wird von der Lehrperson unter Rückgriff auf gängige

---

[3] Zur Unterscheidung von *progress test* vs. *performance test* vgl. auch Bürgi (2007: 60). Mit etwas anderer Gewichtgebung und Terminologie unterscheidet Studer (2004: 76) zwischen "*achievement tests* (Lernfortschritts- oder Sprachstandstests)" und "*proficiency tests* (Leistungs- und Qualifikationsprüfungen)".

[4] In der sozialwissenschaftlichen Forschung wird zu Recht immer wieder betont, dass die beiden Termini normbezogen vs. kriteriumsbezogen eigentlich unglücklich gewählt sind. Jede Messung bezieht sich letztlich auf Kriterien (einen Soll-Zustand), und auch standardisierte Tests entstehen nicht einfach im luftleeren Raum, sondern sind durch Eichung an einer grossen Stichprobe als "Norm" gewonnen worden.

[5] Diesen Unterschied betont Studer (2004: 76), wenn er vom "prospektiven Charakter" von Zertifikatsprüfungen spricht.

Aufgabentypen (Lückentexte, Umformungen, Zusammenfassungen, Fragen zum Text, usw.) speziell für die betreffende Klasse zusammengestellt. Zweck dieses Tests ist zu überprüfen, ob sich die Schüler den in einer bestimmten Zeitperiode vermittelten Stoff angeeignet haben. Die gestellten Aufgaben sind relativ lebensnah und simulieren zum Teil reale Kommunikationssituationen, dies jedoch um den Preis einer erschwerten objektiven Bewertung. Über eine blosse Überprüfung von Kenntnissen hinaus dokumentieren solche Tests auch die Lernfortschritte der Schüler und bilden somit auch eine Grundlage für eine individualisierte Förderung (Perlmann-Balme, 2001: 994ff.)

Demgegenüber bezieht sich ein *standardisierter Test* nicht auf einen zuvor erteilten Unterricht oder auf ein bestimmtes Lehrmittel, sondern er wird kursunabhängig zur Feststellung des zu einem bestimmten Zeitpunkt erreichten globalen Sprachbeherrschungsgrads verwendet. Er kann zu ganz unterschiedlichen Zwecken eingesetzt werden, zum Beispiel als Einstufungstest (mit dem Ziel der Bildung homogener Lernergruppen), als Zulassungsprüfung (z.B. für eine nächsthöhere Ausbildungsstufe), als Sprachenzertifikat (zur Aussenzertifizierung von Sprachkompetenzen) oder als Messinstrument in einer wissenschaftlichen Untersuchung. In der Regel von einem spezialisierten Testinstitut entwickelt, sind standardisierte Tests vergleichsweise valide und objektiv bezüglich Konzeption, Durchführung und Auswertung. Sie erkaufen diesen Vorzug aber oft mit einem relativ niedrigen Grad an Natürlichkeit der zu lösenden Aufgaben. Dies trifft insbesondere für die psychologische und psycholinguistische Forschung zu; bei der standardisierten Sprachstandsmessung kommen nun vermehrt auch "kommunikative" Tests zum Einsatz. Was bei einem standardisierten Test interessiert, ist das globale Resultat (z.B. eine bestimmte erreichte Punktzahl), nicht etwa Phänomene wie der individuelle Lernfortschritt.

Was nun speziell die für wissenschaftliche Zwecke eingesetzten standardisierten Testverfahren von anderen Verwendungszwecken unterscheidet, ist der Umstand, dass sie keinerlei Selektionsfunktion haben, sondern rein diagnostische Ziele verfolgen. Das hat unter anderem zur Folge, dass die Probanden bei deren Absolvierung wahrscheinlich kaum unter Prüfungsstress stehen.

Sowohl in der Schule als auch in der Wissenschaft werden Sprachkompetenzen gemessen. Die eingesetzten Messinstrumente unterscheiden sich jedoch – wie eben festgestellt – stark voneinander. Überspitzt gesagt erzielt man mit wissenschaftlich erprobten Tests relativ objektive Resultate, dies jedoch um den Preis von lebensfremden Aufgaben; für die *teacher-made tests* trifft das Umgekehrte zu. Auch wenn dieses Dilemma teilweise nicht hintergehbar ist, so ist doch zu fragen, ob der Abstand zwischen den beiden Typen von Tests so gross sein muss, wie er heute ist. Das für jeden Einsatz eines Sprachtests idealerweise anzuvisierende Ziel ist natürlich die

Verbindung einer objektiven Messung von Sprachkompetenzen mit Aufgaben, die sich an authentischen Verwendungssituationen orientieren. Auf das hier angesprochene Dilemma soll unter den Schlussbemerkungen zu diesem Beitrag nochmals kurz eingegangen werden.

4. Die im Projekt "zweisprachige Maturität" eingesetzten Tests

Im Rahmen des Nationalen Forschungsprojekts NFP 56 wird gegenwärtig ein von A. Näf und D. Elmiger geleitetes Projekt zur zweisprachigen Matura in der Schweiz durchgeführt[6]. An den beiden Kurzzeitgymnasien *Lycée Jean-Piaget* in Neuchâtel (NE) und *Lycée cantonal* in Porrentruy (JU) wurde unter anderem die Entwicklung der Sprachkompetenz in der Immersionssprache Deutsch der immersiv unterrichteten Klasse (BIL-Klasse) im Rahmen einer Longitudinalstudie (2005-08) über die drei Gymnasialjahre hinweg verfolgt und mit je einer monolingualen Parallelklasse als Kontrollgruppe (REG-Klasse) verglichen.

4.1 Methodologische Vorüberlegungen

Um die Fortschritte bezüglich der Sprachkompetenz zu beobachten, kann man unterschiedlich vorgehen. Man könnte zum einen anhand aller während der drei Jahre angefertigten *classroom-tests* die jeweils erreichten Ergebnisse (in Form von Punktzahlen und/oder Noten) registrieren und statistisch auswerten. Das Ergebnis liesse sich für jeden Schüler der BIL- und REG-Klassen in Form einer Leistungskurve über die drei Jahre hinweg darstellen. Eine entsprechende Kurve der Durchschnittswerte würde dann den Abstand zwischen den beiden Klassentypen veranschaulichen. Zum andern kann man "von aussen" – aus verständlichen Gründen jedoch bloss punktuell – mit standardisierten und/oder informellen Tests den jeweils erreichten Zwischenstand erheben. Um ein wirklich umfassendes Bild des Sprachstands der Schüler einer Klasse zu bekommen, könnte man versuchen, die Innensicht mit der Aussensicht zu kombinieren.

Das erstgenannte Vorgehen musste aus verschiedenen Gründen ausscheiden. Zum einen wegen der umstrittenen Objektivität von Schulnoten im Allgemeinen, zum andern deshalb, weil bei diesem Vorgehen die Leistungen von BIL- und REG-Klassen (diese werden nicht von derselben Lehrperson unterrichtet) sowie diejenigen der beiden Schulen nicht mehr unmittelbar miteinander vergleichbar wären. Wir haben uns deshalb für den Einsatz von kursunabhängigen Leistungstests entschieden, dies mit einer

---

6 Näheres zu diesem Projekt vorläufig unter der NFP-56-Internetseite www.nfp56.ch. Für einen Überblick über die konkrete Ausgestaltung der zweisprachigen Maturität an den 70 Schweizer Gymnasien, welche diesen Ausbildungsgang anbieten, vgl. Elmiger (2008).

Ausnahme. Ebenfalls in den Vergleich einbezogen wurde nämlich die schriftliche Maturitätsprüfung, sozusagen die offizielle (zwar schulinterne, aber an den beiden Gymnasien für alle Parallelklassen eines Jahrgangs identische) Abschlussmessung des erreichten Endzustands[7]. Die Übernahme von deren Resultaten schien uns auch deswegen von Interesse, weil es sich dabei um eine besonders sorgfältig gestaltete und begutachtete Prüfung handelt, die überdies von einem unabhängigen Zweitkorrektor bewertet wird. Hingegen wäre hier ein Vergleich zwischen den beiden Gymnasien (etwa auf der Basis der erzielten Noten) insofern nicht sehr aussagekräftig, als sich die Schlussprüfungen an den beiden Schulen ziemlich stark voneinander unterscheiden, etwa was die getesteten Teilfertigkeiten betrifft[8].

Es stellte sich nun die doppelte Frage, in welchem Zeitabstand die Interventionen stattfinden und welche sprachlichen Teilfertigkeiten dabei gemessen werden sollten. Im Idealfall könnte man sich vorstellen, dass alle sprachlichen Grundfertigkeiten (im Gemeinsamen Europäischen Referenzrahmen werden bekanntlich nicht vier, sondern fünf angesetzt)[9], welche mehrmals während der drei Schuljahre erhoben würden. Realistisch betrachtet musste aber ein derartiges Design zum vornherein ausscheiden. Ein für unsere Forschungsarbeiten wichtiger Grundsatz war, dass der normale Unterrichtsbetrieb durch diese Sprachtests möglichst wenig gestört werden sollte. Es ist denn auch nur unter dieser Voraussetzung, dass uns die Schuldirektoren ihre Einwilligung für ein Forschungsprojekt, das sich über drei Jahre erstreckt, gegeben hatten[10]. Es gilt dabei zu berücksichtigen, dass unsere Interventionen nicht bloss die Sprachtests in den BIL- und REG-Klassen umfassten, sondern auch die Aufnahme von 40 Lektionen von

---

7 An den beiden Gymnasien werden die jeweiligen BIL- und REG-Klassen einer identischen schriftlichen Maturitätsprüfung unterworfen. Man könnte sich natürlich für die BIL-Klassen auch eine anspruchsvollere Prüfung vorstellen. Eine solche Differenzierung wird aber von den dafür Verantwortlichen mit dem Argument abgelehnt, dass dadurch ein mittelmässiger REG-Schüler eine bessere Maturanote erzielen würde als ein objektiv betrachtet leistungsstärkerer BIL-Schüler, der bei einem solchen Szenario infolge des höheren Schwierigkeitsgrads des Tests benachteiligt wäre.

8 So umfasst etwa die Maturitätsprüfung am Lycée Jean-Piaget – im Gegensatz zu jener am Lycée von Porrentruy – einen Subtest zum Hörverständnis; das Umgekehrte gilt für den in Porrentruy praktizierten Übersetzungsteil (*thème*).

9 Die Fertigkeit *Sprechen* wird im GER in die beiden Teilfertigkeiten "an Gesprächen teilnehmen" und "zusammenhängendes Sprechen" aufgeteilt.

10 Es ist hier nun auch der Ort, den beiden Schuldirektoren Mario Castioni (Neuchâtel) und Pierre-Alain Cattin (Porrentruy) für die entsprechenden Bewilligungen unseren verbindlichen Dank abzustatten. Bedanken möchten wir uns auch bei den Lehrpersonen, die uns den Zugang zu ihren Klassenzimmern ermöglicht haben, sowie natürlich bei den beteiligten Schülerinnen und Schülern, welche die Fragebogen und Testformulare bereitwillig ausgefüllt und sich mit der Zeit an unsere Präsenz gewöhnt haben.

immersivem Sachfachunterricht in den Fächern Geschichte, Biologie, Physik, Bildnerisches Gestalten, Philosophie und Sport.

Während Audioaufnahmen von Lektionen den normalen Unterricht kaum stören (Schüler und Lehrpersonen gewöhnen sich relativ schnell daran), greifen die Tests schon stärker in den Schulalltag ein. Da es organisatorisch kaum möglich und überdies den Schülern kaum zumutbar gewesen wäre, die Tests ausserhalb der normalen Schulzeiten anzusetzen, waren wir darauf angewiesen, diese während der regulären Unterrichtsstunden durchzuführen. Eine rein technische Schwierigkeit liegt dabei darin, dass der Fremdsprachenunterricht meist in Form von Einzellektionen von 45 Minuten Dauer erteilt wird.

Es sei hier nur am Rande angemerkt, dass die Schülerinnen und Schüler zu Beginn und am Ende des bilingualen Lehrgangs auch einen Fragebogen zu ihrem Sprachgebrauch und ihren Spracheinstellungen (L1, Deutsch, Englisch) ausgefüllt und eine Selbstevaluation ihrer sprachlichen Teilfertigkeiten (mit Hilfe der Deskriptoren des Europäischen Sprachenportfolio) vorgenommen haben, dies unter anderem mit dem Ziel, sie zu vermehrter Eigenverantwortung und Reflexivität anzuspornen. Auf die einschlägigen Ergebnisse kann jedoch hier nicht eingegangen werden.

4.2 Chronologie der Testinterventionen

Die Tabelle 2 gibt einen Überblick über die Verteilung der Testinterventionen über die drei Jahre hinweg.

C-Test Wortschatz-Test Dez. 2005	C-Test Wortschatz-Test Nov. 2006	Hörverständnis Leseverständnis April 2008	Schriftliche Maturitätsprüfung Juni 2008

Tab. 2: Chronologie der Sprachtest-Interventionen

Bei den BIL- und REG-Klassen wurden im Laufe der Gymnasialzeit insgesamt sechs Sprachtests durchgeführt, und zwar zu Beginn des ersten und des zweiten Schuljahrs sowie gegen Ende des dritten Schuljahrs. Es handelte sich jeweils für beide Gymnasien um die gleichen Tests. Da sich die schriftliche Maturitätsprüfung in beiden Schulen unterscheidet, kann diese nur für den schulinternen Vergleich der BIL- und REG-Klassen ausgewertet werden.

Wir haben uns bei der Sprachstandsmessung für zwei möglichst wenig invasive, zeitökonomische (sowohl bezüglich Einsatz als auch Auswertung) und erprobte Messinstrumente entschieden, einen *C-Test* und einen *Wortschatz-Test* (Wortinitialen-Test). Im dritten Schuljahr kamen ein *Textverständnistest* und ein *Hörverständnistest* des Goethe-Instituts zum Einsatz. Zur Beschreibung und zur Begründung der im Rahmen der vorliegenden Untersuchung verwendeten Tests vgl. Abschnitt 4.3.

Ursprünglich war noch geplant, den initialen Sprachstand der Schüler der beteiligten BIL- und REG-Klassen mit dem (für 14 Sprachen verfügbaren) DIALANG-Test zu bestimmen. Bei DIALANG handelt es sich um einen computergestützten adaptiven Online-Test, der den Testpersonen – je nach den Ergebnissen des vorangeschalteten Einstufungstests – aus einer umfangreichen Item-Bank mit fünf Modulen (Hörverstehen, Grammatik, Schreiben, Wortschatz, Lesen) Aufgaben zuweist, die bezüglich ihres Schwierigkeitsgrads kalibriert sind[11]. Nachdem die – ziemlich aufwändige – Vorbereitung der Informatikräume schon erfolgt war, fiel jedoch der – in Lancaster (GB) beheimatete – Server während mehrerer Monate aus, sodass wir schliesslich auf den Einsatz dieses Einstufungsinstruments verzichten mussten.

Aufwändigere Tests, die alle vier bzw. fünf kommunikativen Sprachfertigkeiten abprüfen, etwa in der Art der Prüfungen des Goethe-Instituts (Zertifikat Deutsch, Goethe Zertifikat B 2) sind sehr zeitintensiv und mussten schon von daher ausscheiden. Im Weiteren hätte die Durchführung von mündlichen Prüfungen (etwa Prüfungsgesprächen in der Art der Goethe-Prüfungen, welche eine reale Kommunikationssituation simulieren) die verfügbaren personellen und finanziellen Ressourcen überstiegen.

4.3 Kurzvorstellung der Tests

Angesichts der beschränkten zur Verfügung stehenden Mittel mussten wir Ausschau halten nach Messinstrumenten mit einem optimalen Verhältnis von Aufwand und Ertrag. Es musste sichergestellt sein, dass zum einen die durchgeführten Tests aussagekräftige Resultate über den Sprachstand der Schüler liefern und dass zum andern durch deren Einsatz der normale Schulbetrieb nicht allzu sehr gestört würde. Es handelt sich hierbei letztlich um eine ethische Frage: Das Ziel des Deutschunterrichts in den vier Klassen war die optimale Vorbereitung der Schüler auf die Maturität, und es durfte natürlich nicht sein, dass wegen unserer Forschungsinterventionen die Erfolgschancen der beteiligten Schüler beeinträchtigt würden. Im Folgenden sollen nun die verwendeten Tests kurz vorgestellt und bezüglich ihrer Stärken und Schwächen charakterisiert werden.

---

[11] Näheres zum Dialang-Test findet sich unter www.dialang.org/german/. Zum *computer-adaptive testing* (CAT) im Allgemeinen vgl. Perlmann-Balme (2001).

4.3.1 C-Test

Der in den Achtzigerjahren von Klein-Braley und Raatz entwickelte C-Test ist zweifellos einer der besterforschten und am häufigsten eingesetzten Sprachtests überhaupt, sodass wir uns hier kurz fassen können. Eine Zusammenstellung der bisherigen Forschungsresultate findet sich bei Grotjahn (2002) und Gogolin (2001) sowie auf den C-Test-Internetseiten.

Methodologisch gesehen beruht der C-Test auf dem Prinzip der reduzierten Redundanz. Als Teststimulus dienen mehrere authentische Lesetexte, bei denen, beginnend mit dem zweiten Satz, jeweils auf rein mechanische Weise die zweite Hälfte jedes zweiten Wortes getilgt wurde. Meist wird ein Test mit fünf inhaltlich unterschiedlichen Kurztexten und insgesamt 100 Lücken konstruiert. Es ist dann die Aufgabe der Probanden, die "beschädigten" Texte unter Rückgriff auf alle ihnen zur Verfügung stehenden Mittel und Strategien zu rekonstruieren.

Der in unserer Untersuchung zum Einsatz gelangte C-Test besteht aus vier inhaltlich unterschiedlichen, dem Alter und Sprachniveau der Probanden angepassten und in sich abgeschlossenen Teiltexten mit ansteigendem Schwierigkeitsgrad. Seine Eignung wurde vorgängig bei einer Anzahl gleichaltriger frankophoner Jugendlicher überprüft. Die für das Ausfüllen des Testformulars (mit 4 x 25 = 100 Lücken) vorgesehene Zeit betrug 20 Minuten.

Nom, prénom _____, classe _____ **Textes avec lacunes à compléter**
Cet exercice se compose de quatre textes courts, chacun portant sur un sujet différent. Chaque texte comporte 25 lacunes. Complétez chaque lacune par la deuxième moitié du mot. La solution correcte doit compter autant de lettres que la première moitié du mot. Quand ce mot est composé d'un nombre impair de lettres, vous devez en ajouter une de plus (voir l'exemple *das Fen**ster***). En complétant, faites attention à la grammaire et au sens du mot (le mot doit s'insérer logiquement dans le contexte).
Exemples: i_ Zürich > i**n** Zürich heute mor___ > heute mor**gen** das Fen___ öffnen > das Fen**ster** öffnen eine Briefma___ > eine Briefma**rke**
Vous avez pour cet exercice environ cinq minutes par texte. Donc, en tout, environ 20 minutes. **Ein Brief** Lieber Max Wie geht es dir? Mir geht es gut. Wir ha___ bald Fer___ und i__ freue mi__ schon dar___. In d__ letzten Woc___ haben w__ an uns____ Schule uns___ Lieblingslehrer gew____. Ich fi___ Frau Geisler, uns___ Englisch- u__ Geschichtslehrerin a_ besten. S__ ist wirk____ eine gu__ Lehrerin. I__ Unterricht i__ interessant, s__ ist ni__ so str___ und s__ kann al___ toll erklären. Ausserdem gibt sie nicht so viele Hausaufgaben auf.

Tab. 3: Erster Subtext des C-Tests (total 100 Lücken)

Auch wenn im Einzelnen nicht recht klar ist, was der C-Test denn nun genau misst, so konnte doch durch eine breite empirische Forschung nachgewiesen werden, dass er die globale Sprachkompetenz (*general language proficiency*) erfasst. Überdies weist er eine hohe Korrelation mit verschiedenen Aussenkriterien auf, zum Beispiel mit Schulnoten, Lehrerurteilen und mit den Ergebnissen anderer Sprachtests wie etwa dem TOEFL[12]. Dass der C-Test mit den beiden *skills* Leseverstehen und Schreiben korreliert, ist nicht weiter erstaunlich. Überraschend ist hingegen, dass er laut neueren Forschungen auch eine hohe Korrelation mit der Teilfertigkeit mündlicher Ausdruck aufweist (Grotjahn, 2002: 9).

Obwohl der C-Test aus kohärenten, authentischen Kurztexten besteht, kann man ihm – wie fast jedem standardisierten Test – eine gewisse Künstlichkeit vorwerfen. Im "realen Leben" ergänzen wir eben keine an ihrem Ende "beschädigten" Wörter. Immerhin kann man sich auch im Alltag durchaus Situationen vorstellen, die mit dieser Aufgabenstellung vergleichbar sind, etwa wenn beim Drucken einer Internet-Seite der rechte Rand und damit das jeweils letzte Wort der Zeile ganz oder teilweise abgeschnitten ist. Ein kompetenter Sprecher ist dabei in aller Regel imstande, die fehlenden Buchstaben und Silben zu rekonstruieren.

4.3.2 Wortschatz-Test

Das Verfügen über einen angemessenen Wortschatz ist für jede Sprachverwendung zentral, ja recht eigentlich die Voraussetzung für den kommunikativen Erfolg beim Einsatz der sprachlichen Teilfertigkeiten. Für die Erhebung des Umfangs des Wortschatzes eines Probanden existieren verschiedene Testmöglichkeiten (vgl. dazu Read, 2000). Für unsere Zwecke haben wir uns für einen schon etwas älteren, unseres Wissens zum ersten Mal von John Macnamara (1969) eingesetzten Test entschieden, der zu den *semantic richness tests* gerechnet wird. Dieser Test geht letztlich auf Wallace E. Lamberts Flüssigkeitstests (*fluency tests*) zurück, welche dieser Pionier der kanadischen Zweisprachigkeitsforschung 1955 zur Bestimmung der dominanten Sprache bei Zweisprachigen entwickelt hat.

Ohne hier auf Einzelheiten der zugrunde liegenden theoretischen Grundannahmen und der Modalitäten der Durchführung und der Auswertung eingehen zu können, sei nur Folgendes angemerkt: Die Aufgabe der Probanden besteht darin, in sehr kurzer Zeit (2 Minuten pro Cluster) so viele

---

[12] Der in den sechziger Jahren in den Vereinigten Staaten entwickelte TOEFL (*Test of English as a Foreign Language*) ist der weltweit am häufigsten abgelegte Sprachtest überhaupt und basierte bis vor kurzem ausschliesslich auf Multiple-Choice-Aufgaben. Es handelt sich um eine einheitliche Sprachprüfung ohne Niveaustufen, bei der es kein Bestehen oder Nichtbestehen gibt, sondern bei der bloss das erreichte Gesamtscore zählt. Seit 2005 ist eine Internet-Version verfügbar, bei der es nun zusätzlich auch noch einen mündlichen Prüfungsteil gibt.

Lexeme wie möglich zu notieren, welche mit einem bestimmten Konsonantenpaar beginnen, zum Beispiel mit *st-, gr-* oder *bl-* (vgl. dazu die Instruktion in Tab. 4). Von grosser Bedeutung ist bei diesem Test die Auswahl der initialen Konsonantenpaare. Es gilt sich vorgängig zu vergewissern, wie viele potentiell in Frage kommende Lexeme im Grundwortschatz für ein bestimmtes Cluster überhaupt existieren. So registriert etwa das in der Westschweiz verwendete *Vocabulaire de base allemand-français* für den Anlaut *Sp-/sp-* insgesamt 25 Lexeme, das teilweise ebenfalls benutzte zweisprachige Schülerwörterbuch von Langenscheidt deren 39[13]. Für dieses Cluster stellen demnach etwa dreissig Lexeme eine Art ideale Obergrenze für die Probanden dar. Dabei ist jedoch von vornherein klar, dass sie diese Zahl natürlich allein schon wegen der knapp bemessenen Zeitvorgabe nicht erreichen können.

Vocabulaire
ALLEMAND
Notez tous les mots allemands commençant par les consonnes ci-dessous qui vous viennent à l'esprit (peu importe que le mot commence par une majuscule ou une minuscule).
Temps à disposition: 2 minutes par question
Exemple allemand:
Sp-/sp-: *Spiel, Spiegel, spät, spontan, sparen, spazieren …*
[...]

Tab. 4: Instruktion zum Wortschatztest (nach Macnamara, 1969; Kolde, 1981)

Dieser informelle Wortschatztest überprüft den Umfang und die Differenziertheit der lexikalischen Kompetenz. Die Hypothese, die ihm zugrunde liegt, lautet: "Je besser jemand eine Sprache beherrscht, desto grösser ist die Anzahl Lexeme, die er in einer vorgegebenen Zeiteinheit zu produzieren imstande ist." Da man jedoch auf den im schwarzen Kasten des Gehirns befindlichen Wortschatzspeicher nur indirekt über konkrete Sprachproduktionen Zugang haben kann, muss man damit rechnen, dass auch hier Performanzfaktoren – wie etwa die momentane Abrufbarkeit eines Lexems – in das Ergebnis mit einfliessen. Es handelt sich dabei aber um eine für die Spracherwerbsforschung grundsätzlich nicht hintergehbare Randbedingung.

Der von uns eingesetzte Test umfasst 4 deutsche und 2 französische Konsonantenpaare (Zeitbedarf für die Testdurchführung inkl. Instruktion: 15 Minuten). Das jeweils erzielte Gesamtscore ist ein gutes Indiz dafür, welches

---

[13] Châtelanat, Ch. & Henzi, Th. (1972): Vocabulaire de base allemand-français. Lausanne; Langenscheidt (1991): Vocabulaire de base allemand. Berlin.

die dominante Sprache der Probanden ist (Macnamara, 1969: 86). Dies ist insbesondere bei jenen SchülerInnen von Interesse, die im Fragebogen Deutsch als ihre Muttersprache angegeben oder sich als zweisprachig bezeichnet haben. Es handelt sich hier um einen Test mit einem sehr guten Verhältnis von Aufwand und Ertrag. Mit geringem Zeitaufwand kann man sich – anhand der als *pars pro toto* fungierenden Lexeme mit bestimmten Anfangskonsonanten – ein Bild vom Gesamtwortschatz machen, über den ein Proband verfügt. Ein weiterer Vorteil dieses kursunabhängig einsetzbaren Vokabular-Tests besteht darin, dass durch diesen nicht Wortschatzkenntnisse in bestimmten Domänen und Sachbereichen zum Tragen kommen, die entweder zufällig vorhanden sind oder im unmittelbar vorangehenden Unterricht erworben wurden (Kolde, 1981: 321f.). Gegen die Verwendung dieses Tests könnte man wiederum vorbringen, dass es sich bei einer solchen nicht durch einen Kontext gestützten Wortproduktions-Aufgabe unter Zeitdruck um eine eher künstliche Übung handelt.

Der Wortschatz-Test wurde von G. Kolde in seiner pionierhaften Studie über die Sprachkontakte in den zweisprachigen Städten Biel / Bienne und Fribourg / Freiburg repliziert. Es zeigte sich dabei, dass die Probanden in ihrer jeweiligen Erstsprache etwa doppelt so viele Items pro Zeiteinheit zu produzieren imstande waren wie in ihrer Zweitsprache (Kolde, 1981: 323).

4.3.3 Hörverständnis- und Leseverständnistest

In mehreren bisherigen Untersuchungen zum immersiven Sachfachunterricht konnte gezeigt werden, dass die Schüler von dieser Unterrichtsform insbesondere hinsichtlich der rezeptiven Fertigkeiten profitieren. So konnte etwa Bredenbröker (2000: 93) nachweisen, dass die Leseverständnis-leistungen der bilingual unterrichteten Schüler schon nach einem Jahr signifikant höher waren als die der Parallelklassen. Der Grund dafür ist wohl, dass die Schüler der BIL-Klassen einer viel stärkeren Exposition in der Immersionssprache (durch die Lehrperson, Lehrmittel, usw.) ausgesetzt sind als jene der Regelklassen[14]. Wie aus den im Rahmen unseres Projekts durchgeführten Audioaufnahmen hervorgeht, ist der Gesamtanteil des mündlichen Outputs der Schüler relativ bescheiden, eine Feststellung, die wohl für den gymnasialen Unterricht im allgemeinen zutreffen dürfte. Während der einzelne Schüler auch im Rahmen des bilingualen Sachfachunterrichts vor allem zuhört und nur wenig Gelegenheit zum Sprechen hat, dürfte

[14] Allerdings kann diese Exposition bei der zweisprachigen Maturität in der Schweiz beträchtlich variieren. Die von den Ausführungsbestimmungen zum MAR 1995 geforderte Minimalzahl von 600 Stunden wird von einigen Gymnasien gerade erfüllt, von andern jedoch um das Doppelte oder gar Dreifache übertroffen (Elmiger, 2008: 32).

demgegenüber das eigene Schreiben, zumindest in Form des Notizen-Machens, etwas stärker zum Zuge kommen.

Um zu verifizieren, ob sich zwischen den BIL- und REG-Klassen unserer Stichprobe ebenfalls ein Leistungsunterschied im Hörverstehen und Leseverstehen nachweisen lässt, werden wir gegen Ende des dritten und letzten Gymnasialjahrs einen einschlägigen Test durchgeführen. Wir greifen dabei auf einen standardisierten Test, nämlich die Zentrale Mittelstufen-prüfung ZMP des Goethe-Instituts zurück. Aus urheberrechtlichen Gründen konnten wir allerdings nicht einen originalen Test einsetzen, sondern einen – bezüglich Art und Schwierigkeitsgrad der Aufgaben jedoch völlig gleichwertigen – Übungssatz (Übungssatz 0.2)[15].

In einer Evaluation des zweisprachigen Ausbildungsgangs am *Lycée des Creusets* in Sitten (VS) konnte Grüninger (2007) für beide Teilfertigkeiten einen signifikanten Unterschied zwischen der dortigen BIL-Maturaklasse und einer monolingualen Parallelklasse nachweisen. Dieser fiel für das Hörver-stehen deutlicher aus als für das Leseverstehen. Bei dieser Studie kam – neben anderen Tests – ebenfalls der ZMP-Übungssatz 0.2 zum Einsatz.

5. Die Resultate der Eingangsmessung

Zum Zeitpunkt der Niederschrift dieses Beitrags (März 2008) ist unsere Longitudinalstudie noch in vollem Gange. Die an der Untersuchung beteiligten Klassen werden ihre Maturitätsprüfung im Juni 2008 ablegen. Es kann hier deshalb noch kein Gesamtüberblick über unsere Sprachstandsmessungen gegeben werden.

5.1 Vorstellung der Ergebnisse

Es soll nun jedoch abschliessend auf das Ergebnis eingegangen werden, welches die Schülerinnen der BIL- und REG-Klassen der beiden Gymnasien zu Beginn ihrer dreijährigen Gymnasialzeit beim C-Test und Wortschatz-Test erzielt haben. Ohne auf Einzelheiten einzugehen, beschränken wir uns auf die Angabe der jeweiligen Klassendurchschnitte und verzichten auf andere statistische Kennwerte wie etwa Streuungsmasse (z.B. Standardabweichung).

Über die Verteilung der an den Sprachtests teilnehmenden Schüler auf die beiden Kantone und die beiden Typen von Klassen orientiert die Tabelle 5. Die Schüler sind zu Beginn ihrer Gymnasialzeit (Herbst 2005) durchschnittlich 15 Jahre alt; 39 von ihnen sind männlichen und 43 weiblichen Geschlechts.

---

[15] Die ZMP-Prüfung ist unterdessen nach den Vorgaben des Gemeinsamen Europäischen Referenzrahmens GER teilweise neu gestaltet und auf die beiden Niveaustufen Goethe Zertifikat B2 und Goethe Zertifikat C1 kalibriert worden.

Kanton	BIL-Klasse	REG-Klasse	Total
JU	22	20	42
NE	20	20	40
Total	42	40	82

Tab. 5: Verteilung der Schüler auf Kantone und Klassentypen

Bei Untersuchungen zum Immersionsunterricht verhält es sich leider oft so, dass man kaum etwas über den biographischen und sprachlichen Hintergrund der Probanden erfährt[16]. Bei unserer – mit einer überschaubaren Zahl von Probanden arbeitenden – Fallstudie wollten wir deshalb auch diesen in seinen Auswirkungen kaum zu überschätzenden Einflussfaktor mit einbeziehen. Alle von uns getesteten Schüler haben einen einschlägigen Fragebogen ausgefüllt und einen Text über ihre Sprachbiographie verfasst, was uns eine differenzierte Analyse der von ihnen in den Sprachtests erzielten Resultate ermöglicht. Neben der Kenntnis anderer Sprachen wie etwa jenen der Immigration interessierte in unserem Zusammenhang natürlich in erster Linie die Art und Intensität eines eventuellen ausserschulischen Bezugs zur Immersionssprache Deutsch. Aufgrund der detaillierten Angaben der Schüler (Familiensprache, Wohnort, usw.) haben wir in der Folge einen Kriterienraster erstellt, um die Anzahl der "Deutschsprachigen" pro Klasse zu ermitteln (de facto müssen diese Schüler jedoch als zweisprachig angesehen werden, absolvieren sie doch den grösseren Teil der Sachfächer auf Französisch). Bei der Auswertung können wir deren Testresultate dort, wo dies sinnvoll erscheint, getrennt halten. Es handelt sich hier keinesfalls um eine nebensächliche Variable. Ob in einer Klasse bloss ein einziger oder aber die Hälfte der Schüler die Immersionssprache auch zu Hause als Familiensprache (Standard oder Schweizerdeutsch) spricht, hat grosse Auswirkungen auf die Resultate. Es sei in diesem Zusammenhang darauf hingewiesen, dass Schüler, welche die Immersionssprache als Erst- und Familiensprache sprechen, in mehreren Kantonen gar nicht zu einem Immersionsprogramm zugelassen werden (Elmiger, 2008: 45).

In den beiden BIL-Klassen haben wir aufgrund des Kriterienrasters 9 Personen als deutschsprachig kategorisiert (JU: 7, NE: 2). Diese werden in den Tabellen 6 und 7 auch gesondert aufgeführt. Da es in den beiden REG-Klassen jedoch nur je eine einzige deutschsprachige Schülerin gibt, soll hier auf eine getrennte Auswertung dieser Variable verzichtet werden.

---

[16] Bei Lys & Gieruc (2005: 55) werden die Schüler mit der Immersionssprache Deutsch als Familiensprache bei der Auswertung auf vorbildliche Weise von den übrigen Schülern getrennt gehalten.

5.1.1 C-Test

Zunächst nun zu den Resultaten des C-Tests. Bei einer maximalen Punktzahl von 100 erzielten die vier Klassen im Durchschnitt (arithmetisches Mittel der Rohwerte aller Schüler) folgende Resultate:

Kanton	BIL-Klasse		REG-Klasse
JU	70.6	Frankophone: 65.5 Deutschsprachige: 81.7	50.8
NE	57.7	Frankophone: 56.8 Deutschsprachige: 62.0	50.2
Mittelwert Total	65.3	Frankophone: 61.4 Deutschsprachige: 77.3	50.5

Tab. 6: C-Test: Klassendurchschnitte der BIL-Klassen und REG-Klassen in den beiden Kantonen

Die Resultate von Tabelle 6 lassen sich zusammenfassend durch die folgenden vier Feststellungen charakterisieren:

- In beiden Kantonen besteht ein grosser Unterschied bezüglich des Gesamtscores zwischen BIL-Klassen und REG-Klassen. Wie ein Vergleich der Mittelwerte mittels eines t-Tests für unabhängige Stichproben ergibt, ist dieser Unterschied in NE signifikant ($t = 2.17$; $p < 0.05$) und in JU hochsignifikant ($t = 4.85$; $p < 0.0001$).
- Die REG-Klassen in JU und NE unterscheiden sich kaum voneinander.
- In der BIL-Klasse in JU erzielen die Deutschsprachigen deutlich höhere Werte als die Frankophonen (81.7 vs. 65.5). Der entsprechende Unterschied ist in NE etwas kleiner (bei allerdings bloss zwei deutschsprachigen Probanden kaum aussagekräftig).
- Die Frankophonen der BIL-Klassen sind deutlich besser als diejenigen der REG-Klassen. Besonders deutlich ist die Mittelwertdifferenz mit fast 15 Punkten (65.6 vs. 50.8) in JU, ein hochsignifikanter Unterschied ($t = 3.81$; $p < 0.001$).

5.1.2 Wortschatz-Test

Die Tabelle 7 enthält die Resultate des Wortschatz-Tests, mit den gleichen unabhängigen Variablen wie in Tabelle 6. Wie oben in 4.3.2 ausgeführt, besteht der Test aus zwei Subtests: einem zur Immersionssprache Deutsch (4 Konsonantengruppen, total 8 Minuten Zeit) und einem zur Erstsprache Französisch (2 Konsonantengruppen, total 4 Minuten Zeit). Pro Treffer

(existierendes Lexem) wurde ein Punkt vergeben[17]. Die Anzahl produzierter Wörter pro Schüler und pro Sprache wurde zusammengezählt und bildet die Rohwerte, auf denen die Klassendurchschnitte von Tabelle 7 beruhen. Die Zahl der Punkte ist theoretisch nach oben offen, de facto ist sie aber nicht nur durch die Sprachkompetenz der Probanden, sondern auch durch die knappe Zeitvorgabe begrenzt. Falls jemand die beiden Sprachen mit gleicher Kompetenz beherrscht, kann man aufgrund des Testdesigns theoretisch unterstellen, dass er in der Immersionssprache Deutsch doppelt so viele Punkte erzielen sollte wie in seiner Erstsprache Französisch. Eine solche Kalkulation hat jedoch bloss den Wert eines Gedankenexperiments.

Im Folgenden gehen wir nur auf diejenigen Resultate ein, die im Wortschatz-Subtest Deutsch erzielt worden sind.

Kanton	BIL-Klasse: Wortschatztest Deutsch			REG-Klasse: Wortschatztest Deutsch
JU	30.0	Frankophone:	26.9	20.7
		Deutschsprachige:	36.7	
NE	23.3	Frankophone:	23.2	22.6
		Deutschsprachige:	24.0	
Mittelwert Total	27.3	Frankophone:	25.1	21.7
		Deutschsprachige:	33.9	

Tab. 7: Wortschatztest Deutsch (Klassendurchschnitt) nach Kanton und Klassentyp

Der Befund von Tabelle 7 lässt sich folgendermassen resümieren:

- In JU besteht ein grosser Unterschied zwischen der BIL-Klasse und der REG-Klasse ($t = 3.14$; $p < 0.01$, sehr signifikant), während in NE die BIL-Klasse bloss geringfügig besser abschneidet.
- Die Differenzen zwischen den REG-Klassen sind nicht sehr gross; NE liegt hier um 2 Punkte höher als JU.
- Für die BIL-Klasse in JU zeigt die Feinanalyse, dass das hohe Gesamtscore in erster Linie auf das Konto der Deutschsprachigen zu setzen ist.
- Die Frankophonen der BIL-Klasse in JU sind signifikant besser als diejenigen der REG-Klassen ($t = 2.38$; $p < 0.05$); in NE ist der Unterschied nur geringfügig.

[17] Die objektive Auswertung dieses Tests bedarf mehrerer zusätzlicher Konventionen. So wurden etwa leichte orthographische Verstösse toleriert. In den sehr seltenen Fällen, wo jemand statt eines Lexems zwei Wortformen desselben Lexems notierte (*sprechen*, *sprach*), wurde nur ein Punkt dafür vergeben.

5.2 Interpretation der Ergebnisse

Wenn wir die Ergebnisse beider Tests zusammenfassend in den Blick nehmen, so kann man sagen, dass die BIL-Klassen ganz generell bessere Sprachleistungen erbringen als die REG-Klassen. Diese Aussage behält auch dann ihre Gültigkeit, wenn man bei den BIL-Klassen nur die Werte der Französischsprachigen ("echt Frankophonen") in Betracht zieht[18]. Im Weiteren ist von Bedeutung, dass die Ergebnisse der beiden Tests in hohem Grade konsistent sind und miteinander korrelieren. Mit andern Worten: Wer im C-Test eine hohe Punktzahl erzielt, schneidet auch im Wortschatz-Test gut ab. So ist etwa der Rangkorrelations-Koeffizient (nach Spearman) für die BIL-Klasse in JU auf dem 1% Niveau signifikant (R = 0.67). Dieses Ergebnis ist nicht zuletzt in Anbetracht der beiden völlig verschiedenartigen Testformate bemerkenswert: Beim integrativen C-Test handelt es sich um eine gebundene, vollständig vorstrukturierte Aufgabenstellung, beim "sektoriellen" Wortschatz-Test dagegen um ein kreatives freies Assoziieren anhand eines Stimulus.

Die hier vorgestellte Sprachstandsmessung erfolgte im ersten Schuljahr zu Beginn des immersiven Ausbildungsgangs. Das bedeutet nun aber, dass die höhere Sprachkompetenz der BIL-Klassen nicht eine Auswirkung des Immersionsunterrichts sein kann! Vielmehr ist es so, dass die Schüler dieser Klassen schon bei ihrem Eintritt ins dreijährige Gymnasium in Deutsch besser waren als jene der monolingualen Parallelklassen[19]. Wenn man gravierende Fehlschlüsse vermeiden will, gilt es, bei der Interpretation der Schlussresultate diesen keineswegs trivialen Umstand im Auge zu behalten. Es zeigt sich hier mit aller Deutlichkeit, dass es für die Abklärung der Wirkung von immersivem Unterricht nicht bloss punktuelle Messungen, sondern auch und vor allem Longitudinalstudien braucht (oder aber, falls dies nicht möglich ist, mindestens Pseudo-Longitudinalstudien). Denn wenn man durch eine Momentaufnahme der Sprachkompetenz, z.B. im letzten Gymnasialjahr, zum Ergebnis kommt, dass die BIL-Klassen den Parallelklassen um so und so viel überlegen sind, so bedeutet das noch nicht, dass diese Differenz ganz oder teilweise auf den bilingualen Ausbildungsgang zurückzuführen ist. Vieles deutet nämlich darauf hin, dass Immersionsprogramme die überdurchschnittlich motivierten und leistungsstarken Gymnasiasten anziehen. Schon allein die Tatsache, dass an vielen Gymnasien spezielle Aufnahmebedingungen und Selektionshürden

---

[18] Was die Unterschiede zwischen den Kantonen betrifft, so sollte man das bessere Abschneiden der Frankophonen der BIL-Klasse in JU wohl nicht überbewerten; nur durch die Untersuchung der Studienanfänger über mehrere Jahrgänge hinweg könnte man hier zu zuverlässigen Aussagen gelangen.

[19] Eine analoge Feststellung macht Bredenbröker (2002: 93) für den Beginn der Englisch-Immersion im Bundesland Niedersachsen.

bestehen[20], deutet darauf hin, dass wir es hier mit einer Art von Begabtenförderung zu tun haben, oder vielleicht eher mit einer Art "Elitenbildung durch die Hintertür" (vgl. dazu die Einführung von A. Näf in: Elmiger, 2008: 6).

5.3 Zum Kompetenzniveau der zweisprachigen Maturität

In jüngster Zeit ist in Kreisen der Bildungsverantwortlichen eine Diskussion darüber entstanden, was denn der Vermerk "zweisprachige Maturität" im Maturitätszeugnis genau bescheinige. Im Gegensatz zu einem internationalen Sprachenzertifikat wird damit nämlich nicht bestätigt, dass die Absolventen ein bestimmtes Kompetenzniveau erreicht haben. Der Eintrag attestiert lediglich, dass einzelne Fächer nicht in der lokalen Schulsprache, sondern in einer Immersionssprache (Landessprache oder Englisch) unterrichtet und geprüft worden sind. Um diesem Unterricht folgen zu können, müssen die Teilnehmer wenn möglich schon von Anfang an über eine – so die gängige Redensweise – "funktionale Zweisprachigkeit" verfügen. Dieser nur schwer fassbare Terminus kann jedoch vieles in sich schliessen, und er sagt letztlich nicht sehr viel über den tatsächlichen Grad der Sprachbeherrschung aus.

Auch wenn somit bis heute noch weitgehend Unklarheit darüber herrscht, was man über den durch die zweisprachige Maturität erreichbaren Sprachstand aussagen kann beziehungsweise was man als realistisches Ziel anvisieren sollte[21], so ist hingegen klar, was *nicht* deren Ziel sein kann, nämlich eine muttersprachnahe Beherrschung (*native-like L2 proficiency*) der Immersionssprache[22]. Dafür ist eine Exposition während der drei bis vier Gymnasialjahre nach der Formel "späte Teilimmersion" einfach nicht ausreichend. Dies gilt auch für jene Schulen, in denen die "Gesamtdosis" zweimal oder gar dreimal höher ist als die für die eidgenössische Anerkennung geforderte Mindestzahl von 600 Stunden (vgl. dazu Elmiger, 2008: 31f.). Ferner weiss man bis heute kaum etwas darüber, inwieweit die zum Teil doch eher restringierte

---

[20] An einem der beiden von uns untersuchten Gymnasien existieren keine explizit kodifizierten Aufnahmebedingungen, am andern müssen der allgemeine Notendurchschnitt und die Deutschnote mindestens genügend sein.

[21] Eine realistische Vorgabe – in den Standards des Gemeinsamen Europäischen Referenzrahmens ausgedrückt – könnte etwa das Niveau B 2 für die produktiven und C 1 für die rezeptiven Fertigkeiten sein. Die Frage ist dann allerdings, vom wem das Erreichen eines solchen Kompetenzniveaus bescheinigt werden sollte. Ferner, falls eine Aussenzertifizierung gewählt würde (gemäss dem Grundsatz dass nicht prüft, wer ausbildet), was dies für die öffentlichen Gymnasien für rechtliche, organisatorische und finanzielle Konsequenzen hätte.

[22] Dass eine muttersprachnahe L2-Kompetenz nur für eine ganz kleine Zahl von "späten" L2-Lernern mit hoher Sprachbegabung und unter besonders günstigen Umständen anvisierbar ist, wird in der Forschungsliteratur immer wieder betont. Bei Geiger-Jaillet (2005: 28) werden nicht weniger als elf den L2-Erwerb positiv beeinflussende Faktoren genannt. Im "Volksmund" wird oft akzentfreie Aussprache in der L2 mit muttersprachnaher Sprachkompetenz gleichgesetzt.

Fachsprache bestimmter Disziplinen wie etwa der Mathematik[23] sprachliche Mittel bereitstellt, die mit Gewinn auf die allgemeine Sprachkompetenz transferierbar sind.

Wie durch zahlreiche Forschungen nachgewiesen werden konnte, vereinigt eine zweisprachige Person nicht einfach additiv die Komptenzen zweier einsprachiger Individuen in sich. Vielmehr handelt es sich dabei um eine domänenspezifisch geregelte Gesamtkompetenz *sui generis*. Eine ausgeglichene Zweisprachigkeit (*balanced bilingualism*) ist denn auch eher ein theoretisches Konstrukt als eine bei konkreten Individuen nachweisbare Realität.

Wenn also muttersprachnahe Kompetenz nicht das Ziel immersiven Unterrichts sein kann[24], dann stellt die Sprachkompetenz von *native speakers* natürlich auch keine sinnvolle Messlatte für den durch die zweisprachige Maturität erreichbaren Schlussstand dar. Bei einem solchen Vergleichsmassstab wäre der Blick zu sehr auf die noch vorhandenen Defizite (z.B. bezüglich sprachlicher Korrektheit) fixiert statt auf die trotz Kenntnislücken schon assimilierten Sprachmittel und kommunikativen Strategien. Eine sinnvollere Referenzgrösse wären dagegen Erwachsene, die in Alltag und Beruf eine funktionale Zweisprachigkeit praktizieren[25]. Wie dies auch in anderen Studien der Fall ist, haben jedoch auch wir uns hier dafür entschieden, als Vergleichsgrösse die gleichaltrigen Gymnasiasten einer "monolingualen" Parallelklasse heranzuziehen und die Differenz zwischen diesen beiden Gruppen zu messen und zu analysieren.

In der Fachliteratur zum C-Test wird davon ausgegangen, dass ein Muttersprachler[26] im Prinzip imstande sein sollte, die Lücken zu 100% Prozent zu ergänzen (Grotjahn, 2002: 216). Um zu prüfen, ob der von uns eingesetzte C-Test diese Anforderung erfüllt, haben wir diesen im Frühjahr 2006 von einer

---

23 Wie die Fragebogenerhebung von Elmiger (2008: 30) zutage gefördert hat, sind an den öffentlichen Gymnasien der Schweiz Geschichte und Mathematik die mit Abstand am häufigsten immersiv unterrichteten Fächer.

24 Im Informations- und Werbematerial mehrerer Schweizerischer Gymnasien wird ausdrücklich betont, dass der immersive Ausbildungsgang keineswegs den Anspruch erhebt, "perfekt zweisprachige" Personen heranzubilden. Eine solche Klarstellung ist deshalb wichtig, weil die Schulleitungen zum Teil von der Elternschaft her mit eher unrealistischen Erwartungen konfrontiert werden.

25 Was funktionale Zweisprachigkeit konkret bedeutet, kann man etwa anhand von (mehr oder weniger spontanen) Radio- und Fernsehinterviews von Bundesräten in der jeweils anderen Landessprache anschaulich aufzeigen und untersuchen.

26 Selbstverständlich handelt es sich auch bei *Muttersprachler* oder *native speaker* um ein idealisiertes theoretisches Konstrukt, das ebenfalls eine relativ breite Palette von Realitäten abdeckt. Klar ist aber, dass Muttersprachler nicht bloss viel weniger, sondern vor allem völlig andere Normverstösse begehen als L2-Lerner der betreffenden Sprache (z.B. kaum Genusfehler).

Klasse gleichaltriger Gymnasiasten an der Kantonsschule Zug (ZG) ausfüllen lassen. Das Ergebnis: Der von den Zuger Probanden erzielte Klassendurchschnitt beträgt 95.1 Punkte und kommt damit tatsächlich dem Richtwert von 100% sehr nahe. Neben diesem Globalresultat sind hier nun aber zwei weitere Beobachtungen von grossem Interesse. Zum einen sind die Zuger Resultate, was die Streuung der Punktzahlen der Teilnehmer betrifft, sehr homogen, dies im Gegensatz zu den Westschweizer Klassen, bei denen sowohl in den BIL- als auch in den REG-Klassen zwischen stärkeren und schwächeren Schülern enorme Unterschiede bestehen. Zum andern spielt die Variable des ansteigenden Schwierigkeitsgrads der vier Teilaufgaben bei Muttersprachlern offensichtlich keine Rolle. Denn während bei den Westschweizern tatsächlich die erste Aufgabe am besten und die vierte am schlechtesten gelöst wurde, haben die Zuger Gymnasiasten alle Teilaufgaben annähernd gleich gut bewältigt.

Und noch etwas wird durch diesen Vergleich deutlich. Der Durchschnittswert von ZG liegt mit 95.1 Punkten auch weit über jenem der Deutschsprachigen der BIL-Klassen in JU (81.7) und NE (62.0). Dies ist nicht weiter erstaunlich: Die "Deutschsprachigen" unserer Stichprobe sprechen zwar in der Familie exklusiv oder teilweise deutsch bzw. schweizerdeutsch; sie haben jedoch ihre obligatorische Schulzeit auf Französisch durchlaufen und ihre Freizeit überwiegend mit frankophonen Gleichaltrigen verbracht.

Alles in allem kann man behaupten, dass der C-Test nicht nur mit grosser Zuverlässigkeit Muttersprachler von Nicht-Muttersprachlern unterscheiden kann, sondern sich auch für die vergleichende Feststellung des Grads der Zweisprachigkeit eignet. Zweisprachigkeit ist eben letztlich nichts anderes als ein bestimmter – je nach den jeweiligen Bedürfnissen festgesetzter – Fixpunkt auf einem Kontinuum.

Wie in 4.3.2 ausgeführt, wurde der von uns eingesetzte Wortschatz-Test ursprünglich zur Bestimmung der dominanten Sprache bei Zweisprachigen entwickelt. Wir haben auch diesen mit den Zuger Gymnasiasten durchgeführt. Ohne auf Einzelheiten einzugehen, kann man sagen, dass die Zuger Probanden beim Subtest Deutsch in der gleichen Zeitspanne rund doppelt so viele Lexeme zu produzieren imstande waren wie die Westschweizer BIL-Klassen und fast dreimal so viele wie die REG-Klassen (Klassendurchschnitt ZG: 56.8 Lexeme, vgl. dazu die Werte in Tab. 7). Beim französischen Teil des Wortschatztests fällt dagegen das Score der Zuger Gymnasiasten ziemlich bescheiden aus (Durchschnitt 13.1 Lexeme), jedenfalls wesentlich unter dem von den Westschweizern erzielten Klassendurchschnitt in ihrer L2 Deutsch. Alles in allem kann man behaupten, dass die Fähigkeit, eine Menge von Wörtern mit einer bestimmten initialen Konsonantengruppe zu produzieren, auf enge Weise mit der Variablen korreliert, ob es sich dabei um die

Muttersprache (bzw. die dominante Sprache) oder aber um eine Fremdsprache handelt, etwa eine durch gesteuerten Unterricht vermittelte L2.

6. Einbettung in den bildungspolitischen Kontext

Sprachkompetenzen messen in Schule und Wissenschaft, so lautet der Titel unseres Beitrags. Wie wir oben aufgezeigt haben, gibt es gute Gründe dafür, in beiden Bereichen unterschiedliche Testverfahren einzusetzen. Aber letztlich geht es beiderorts eben doch um das Gleiche, nämlich um die Feststellung von sprachlichen Kompetenzen. Kein Zweifel: Die Ansprüche an die Zuverlässigkeit, Transparenz und Vergleichbarkeit von Prüfungen, Tests und Zensuren sind in den letzten Jahren stark gestiegen. Harmonisierung, Schulkoordination, Nationale Treffpunkte, Bildungsmonitoring, Standardisierung, Output-Orientierung, Mindestanforderungen und Qualitätsentwicklung: dies nur eine Auswahl von programmatischen Fahnenwörtern, die gegenwärtig in aller Munde sind. Zweifellos hat die Schweiz mit ihren 26 kantonalen Bildungsdirektionen diesbezüglich im europäischen Vergleich einen gewissen Nachholbedarf. Für die obligatorische Schulzeit werden gegenwärtig im Rahmen des Harmos-Konkordats für die schulischen Kernfächer an den sog. Schnittstellen am Ende des 2., 6. und 9. Schuljahrs Minimalstandards entwickelt.

Diese neuerdings in Gang gekommene Dynamik der Harmonisierung und Standardisierung dürfte auch vor den Toren des Gymnasiums nicht Halt machen. Im Maturitäts-Anerkennungs-Reglement MAR von 1995 (teilrevidiert 2007) wird zwar die Grobstruktur des gymnasialen Ausbildungsgangs (Dauer, Fächer, Prüfungsmodalitäten, usw.) festgeschrieben. Die Konkretisierung dieses allgemeinen Rahmens durch Lehrpläne wird jedoch den Kantonen überlassen, und was die inhaltliche Füllung dieser Rahmenbestimmungen betrifft, herrscht für die Schulen, ja für jede einzelne Lehrperson eine grosse Freiheit. Damit sind die Gymnasien bisher zwar nicht schlecht gefahren. Aber in einer Zeit, in der auf der obligatorischen Schulstufe nicht bloss interkantonale Vergleiche, sondern Länder-Hitparaden wie diejenigen von PISA die öffentliche Diskussion beherrschen, wäre es naiv zu glauben, dass diese Welle nicht auch auf die Sekundarstufe II überschwappen wird. Über eine gewisse Standardisierung in den Kernfächern wird insbesondere für den Anfang und das Ende des gymnasialen Ausbildungsgangs nachgedacht, z.B. über identische Übertrittsprüfungen pro Kanton (seit 2007 an den Langzeitgymnasien des Kantons Zürich realisiert) oder über identische Maturitätsprüfungen pro Schule oder gar pro Kanton[27].

---

[27] Dagegen sind Szenarien wie eine einheitliche sprachregionale oder gar nationale Maturität in der föderalistischen Schweiz zumindest bislang kein Thema, dies im Gegensatz zu

Es macht den Anschein, dass bei der Diskussion um Bildungsstandards dem Fremdsprachenunterricht eine Vorreiterrolle zukommt, dies wohl deshalb, weil dieser mit den einzelsprachunabhängigen Niveaubeschreibungen des GER über eine europaweit diskutierte und vielerorts bereits auch zum Einsatz gelangende "Messlatte" verfügt. Deren Befürworter betonen, dass Bildungsstandards zu einer transparenteren und damit auch gerechteren Aufgaben-, Test- und Beurteilungskultur beitragen und die bisher herrschende Praxis positiv beeinflussen können. Allerdings artikuliert sich unterdessen auch eine gewisse Skepsis gegenüber allfälligen unerwünschten Nebenfolgen. Befürchtet wird insbesondere die Rückwirkung von Minimal- oder Regelstandards auf den gesamten Unterricht, der sog. *Washback*-Effekt. Zwischen Lehrpersonen und Bildungsverantwortlichen besteht indes weitgehend Einigkeit darüber, dass eine allfällige Einführung von Bildungsstandards an den Schweizer Gymnasien nicht zu einem durchgehenden *teaching the test* führen darf, während gegen ein in den Unterricht eingebettetes *teaching to the test,* das bereits bisher zum Schulalltag gehört, nichts einzuwenden ist[28].

Welchen Weg das Schweizer Gymnasium mittelfristig auch immer einschlagen wird: Prüfungen werden mit Sicherheit auch in Zukunft zum Alltagsgeschäft gehören. Dabei werden wohl weiterhin überwiegend informelle Tests zum Einsatz kommen. Ob darüber hinaus an bestimmten "Gelenkstellen" auch standardisierte Tests eingeplant werden, ist eine politische, keine wissenschaftliche Entscheidung. Hingegen scheint weitgehend Konsens darüber zu herrschen, dass eine "Amerikanisierung" der Schweizer Schulen vermieden werden soll. In den Vereinigten Staaten herrscht bekanntlich ein Misstrauen gegenüber intuitiv korrigierten Prüfungsaufgaben, und es kommen deshalb in weitem Umfang geschlossene Aufgabenformate wie die korrekturfreundlichen *Multiple Choice*-Tests zum Einsatz. Entsprechend hochentwickelt ist denn dort auch die Theorie und Praxis der Distraktorenformulierung.

Wie eingangs festgestellt, besteht zwischen dem Testen in der Schule und jenem in der Wissenschaft ein eigentlicher Graben. Wir wollen nun aber abschliessend doch noch die Frage in den Raum stellen, ob dieser wirklich so tief sein müsste, wie er heute ist, und ob nicht die beiden Domänen Schule

---

Deutschland, wo nach dem für jeweils ein Bundesland gültigen Zentralabitur jetzt auch eine Diskussion über ein bundesweites Einheitsabitur eingesetzt hat. Standardisiert und sogar weltweit identisch ist im übrigen das *International Baccalaureate IB*, das auch in der Schweiz immer häufiger abgelegt wird, jedoch nur unter besonderen Bedingungen zur Immatrikulation an einer Schweizer Universität berechtigt.

[28] Eine solches bloss den abprüfbaren Stoff der Abschlussexamen trainierendes *teaching the test* wird in der Schweiz gelegentlich den Privatschulen vorgeworfen, welche auf die Schweizerische (früher: Eidgenössische) Maturität vorbereiten.

und Wissenschaft voneinander lernen könnten. Die Wissenschaft könnte sich bei der Testkonstruktion von den die reale Sprachverwendung simulierenden Aufgaben der schulischen Prüfungsformen inspirieren lassen, während umgekehrt die Schule stärkere Anstrengungen in Richtung Validität und Objektivität der im Unterricht eingesetzten Prüfungen unternehmen müsste. Dabei könnten auch standardisierte Prüfungen wie die internationalen Sprachenzertifikate ihren Beitrag leisten. Nicht indem sie schulinterne Prüfungen ersetzen, sondern indem sie diese auf sinnvolle Weise ergänzen[29]. Es gilt dabei aber stets im Auge zu behalten, dass der Fremdsprachenunterricht auch und gerade am Gymnasium nicht auf rein sprachpraktische Ziele reduziert werden darf[30].

Alles in allem: Die Schweizer Gymnasien sind gut beraten, wenn sie in Zukunft auf eine Mischung von informellen und standardisierten Tests setzen, um so die Stärken beider nutzen zu können und um damit dem Idealziel etwas näher zu kommen: der objektiven Messung von Sprachkompetenzen anhand von möglichst wirklichkeitsnahen Aufgabestellungen.

BIBLIOGRAPHIE

Beck, B. & Klieme, E. (Hg.) (2007): Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim / Basel (Beltz Verlag).

Bredenbröker, W. (2000): Förderung der fremdsprachlichen Kompetenz durch bilingualen Unterricht. Empirische Untersuchungen. Frankfurt a. M. (Lang).

Bürgi, H. (2007): Im Sprachbad. Besseres Englisch durch Immersion: eine Evaluation zweisprachiger Ausbildungsgänge an drei kantonalen Gymnasien in der Schweiz. Bern (h.e.p.).

Elmiger, D. (2008): Die zweisprachige Maturität in der Schweiz. Die variantenreiche Umsetzung einer bildungspolitischen Innovation. Mit einer Einführung von Anton Näf. Bern (Staatssekretariat für Bildung von Forschung SBF).

Europarat (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Hg. vom Goethe-Institut, der KMK, der EDK und dem BMBWK. Berlin (Langenscheidt).

Geiger-Jaillet, A. (2005): Le bilinguisme pour grandir. Naître bilingue ou le devenir par l'école. Paris (L'Harmattan).

Gogolin, I. (2001): Sprachstandsdiagnosen. In: G. Helbig, L. Götze, G. Henrici & H.-J. Krumm (Hg.): Deutsch als Fremdsprache. Ein internationales Handbuch. Band 2, Berlin (de Gruyter), 1007-1016.

Grotjahn, R. (Hg.) (2002): Der C-Test. Theoretische Grundlagen und praktische Anwendungen, Band 4, Bochum (AKS-Verlag).

---

29 Zu dieser Konklusion gelangt auch Thomas Studer (2004: 94) in seinem Beitrag über die Internationalen Sprachenzertifikate im Bereich DaF in der Schweiz.

30 Vor dieser Gefahr warnt etwa Konrad Schröder in einem Beitrag zur DESI-Studie (Deutsch Englisch Schülerleistungen International) (Beck & Klieme, 2007: 294)

Grüninger, S. (2007): Der zweisprachige Studiengang im Lycée-Collège des Creusets im Kanton Wallis. Lizentiatsarbeit Universität Neuchâtel.

Hollenweger, J. *et al.* (2005): Schlussbericht Evaluation "Zweisprachiger Ausbildungsgang an Mittelschulen". Zürich (Pädagogisches Institut der Universität Zürich).

Kühn, I., Lehker, M. & Timmermann, W. (Hg.) (2005): Sprachtests in der Diskussion. Frankfurt a. M. (Peter Lang).

Lys, I. & Gieruc, G. (2005): Etude de la maturité bilingue dans le canton de Vaud. Enjeux, outils d'évaluation et niveaux de compétence. Lausanne (URSP).

Macnamara, J. (1969): How can one measure the extent of a person's bilingual proficiency? In: L. G. Kelly (ed.), Description and Measurement of Bilingualism. Toronto (University of Toronto Press), 80-97.

McNamara, T. F. (1996): Measuring second language performance. London (Longman).

Morfeld, P. (2003): Sprachenzertifikate. In: K.-R. Bausch, H. Christ & H.-J. Krumm (Hg.): Handbuch Fremdsprachenunterricht, 4. Auflage, Tübingen (A. Francke Verlag), 384-387.

Perlmann-Balme, M. (2001): Leistungsmessung. In: G. Helbig, L. Götze, G. Henrici & H.-J. Krumm (Hg.): Deutsch als Fremdsprache. Ein internationales Handbuch. Band 2, Berlin (de Gruyter), 994-1006.

Read, J. (2000): Assessing vocabulary. Cambridge (Cambridge University Press).

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

Studer, Th. (2004): Internationale Zertifikate für Deutsch als Fremdsprache in der Schweiz. In: Th. Studer & G. Schneider (Hg.): Deutsch als Fremdsprache und Deutsch als Zweitsprache in der Schweiz, Bulletin Suisse de Linguistique Appliquée, 79, 69-97.

Vollmer, H. J. (2003): Leistungsmessung, Lernerfolgskontrolle und Selbstkontrolle. In: K.-R. Bausch, H. Christ & H.-J. Krumm (Hg.): Handbuch Fremdsprachenunterricht, 4. Auflage, Tübingen (A. Francke Verlag), 365-370.