Schneider, Günther

Auf dem Weg zu Skalen für die rezeptiven Kompetenzen im Bereich des Englischen

Beck, Bärbel [Hrsg.]; Klieme, Eckhard [Hrsg.]: Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistung International). Weinheim u.a. : Beltz 2007, S. 273-289

urn:nbn:de:0111-opus-32363

in Kooperation mit:

BELTZ

http://www.beltz.de

Bärbel Beck / Eckhard Klieme (Hrsg.)

Sprachliche Kompetenzen

Konzepte und Messung

DESI-Studie
(Deutsch Englisch Schülerleistungen International)

Beltz Verlag · Weinheim und Basel

Dr. *Bärbel Beck* ist Diplompsychologin und Projektkoordinatorin am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt a.M.

Prof. Dr. *Eckhard Klieme* ist Direktor des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) in Frankfurt a.M.

Diese Studie wurde im Auftrag der Kultusministerkonferenz erstellt. Für die Richtigkeit des Ergebnisses der Studie trägt das »Deutsche Institut für Internationale Pädagogische Forschung« allein die Verantwortung.

Lektorat: Peter E. Kalb

© 2007 Beltz Verlag · Weinheim und Basel
www.beltz.de
Herstellung: Klaus Kaltenberg
Satz: Deutsches Institut für Internationale Pädagogische Forschung
Druck: Druckhaus »Thomas Müntzer«, Bad Langensalza
Printed in Germany

ISBN 978-3-407-25398-9

Inhaltsverzeichnis

Günther Schneider

Auf dem Weg zu Skalen für die rezeptiven Kompetenzen im Bereich des Englischen

In den Bildungsstandards der deutschen Kultusministerkonferenz für die erste Fremdsprache (KMK 2004) wurde der Weg gewählt, die Standards von Expertengruppen durch eine Anpassung und Ergänzung von Kompetenzbeschreibungen des „Gemeinsamen europäischen Referenzrahmens für Sprachen" (Europarat 2001) formulieren zu lassen und die verbalen Kompetenzbeschreibungen durch Testbeispiele zu illustrieren. Dabei handelt es sich sowohl um Übernahmen aus bestehenden internationalen Sprachdiplomprüfungen als auch um Aufgaben aus Lehrwerken und neu entwickelten Aufgaben.

Für die Teile aus Diplomprüfungen ist festzuhalten, dass die vorgenommene Zuordnung zu einem bestimmten Referenzniveau zunächst nur eine Hypothese ist. Denn bisher beruhen solche Zuordnungen noch nicht auf kontrollierten Verfahren, wie sie das Handbuch des Europarats für die Zuordnung von Prüfungen zum Referenzrahmen vorsieht (Council of Europe 2003). Bei den aus Lehrwerken zusammengestellten Aufgaben und den „neuen" Aufgaben im KMK-Entwurf erhält man den Eindruck, dass die Aufgabenbeispiele allzu schnell zusammengestellt wurden. In den einleitenden Bemerkungen zu den Aufgabenbeispielen heißt es: „Hinsichtlich der Orientierung am Gemeinsamen europäischen Referenzrahmen und ggf. auch im Lichte der Erfahrungen von künftigen Vergleichsuntersuchungen sind sie (die Aufgaben) zu überprüfen und weiter zu entwickeln" (KMK 2004, S. 22-23).

Mit DESI wurde ein anderer, wissenschaftlich abgestützter Weg beschritten, indem theoriebasiert und durch empirische Analyse von Aufgaben Kompetenzmodelle und Niveaubeschreibungen entwickelt wurden, die dann generalisierend mit den Niveaubeschreibungen des Referenzrahmens verknüpft werden.

Zu den großen Vorteilen des Projekts gehört u.a., dass die Entwicklung der Kompetenzmodelle für verschiedene Fertigkeitsbereiche auf einem gemeinsamen Konzept und einer gemeinsamen Methodologie beruht, dass beträchtliche Ressourcen vorhanden waren und eine sehr große Stichprobe zur Verfügung stand. Sowohl die verwendete Entwicklungsmethode wie die Ergebnisse sind potentiell relevant für die Konkretisierung vorliegender Bildungsstandards und auch für die Entwicklung von Bildungsstandards in anderen Ländern.

So hat z.B. die Schweizerische Konferenz der kantonalen Erziehungsdirektoren den Auftrag erteilt, im Rahmen eines Projekts zur Harmonisierung der obligatorischen Schule (HarmoS) u.a. für die Muttersprachen und die Fremdsprachen, die in den verschiedenen Sprachregionen der Schweiz Unterrichtsfächer sind, Bildungsstandards auszuarbeiten und dazu ein Kompetenzmodell mit verschiedenen Kompetenzniveaus

zu entwickeln, wobei der „Gemeinsame europäische Referenzrahmen für Sprachen“ als Grundlage dienen soll (Schneider 2005).

Solche nachfolgenden Projekte sind in der günstigen Lage, auf den Erfahrungen, die im DESI-Projekt gemacht wurden, aufbauen zu können. Aus dieser Optik empfiehlt sich eine kritische Auseinandersetzung mit den Methoden und Resultaten des DESI-Projekts. Ich gehe auf folgende Aspekte des DESI-Projekts im Bereich des Englischen ein und konzentriere mich auf den Bereich der rezeptiven Kompetenzen:

- das Testkonstrukt und die DESI-Testaufgaben im Vergleich zu Tests für internationale Sprachdiplome;
- die schwierigkeitsbestimmenden Merkmale;
- die Formulierung der Skalen und ihre Beziehung zum Referenzrahmen.

Gerade in Bezug auf die rezeptiven Kompetenzen sind die Erwartungen an das DESI-Projekt groß. Denn die Kompetenzmodelle zum Hör- und Leseverstehen sind im Referenzrahmen, wie verschiedentlich kritisiert wurde, zu wenig ausgebaut (z.B. Alderson u.a. 2004; Vollmer 2003). Anders als im Literacy-Modell von PISA werden kognitive Prozesse im Kompetenzmodell des Referenzrahmens zwar mitbedacht, sind dort aber nicht stufenbildend.

Ich beziehe mich auf die Arbeiten zu den rezeptiven Kompetenzen, die an der DESI-Fachtagung 2004 in der Arbeitsgruppe für den Bereich des Englischen vorgestellt wurden, und auf die Beiträge von Nold/Rossa, Nold/Willenberg und Harsch/Schröder in diesem Band. Ich greife dabei auch kritische Einwände auf, die in den Diskussionen an der Fachtagung vorgebracht wurden.

Die DESI-Tests in der Testlandschaft

In internationalen Sprachdiplomprüfungen (z.B. Cambridge-Prüfungen, IELTS, dem neuen TOEFL-Test) ist es inzwischen selbstverständlich, dass Tests zum Lese- und Hörverstehen mehrere Subtests, Textsorten und Aufgabenformate enthalten. Deutlich erkennbar ist dabei das Bemühen, bei der Gestaltung der Aufgaben realen Anwendungssituationen nahe zu kommen und deren relevante Merkmale zu berücksichtigen. Dazu gehört z.B., dass versucht wird, durch die Angabe einer Lese- bzw. Hörsituation entsprechende Lese- bzw. Hörintentionen zu stimulieren (oder wenigstens zu simulieren) und Fertigkeiten auch kombiniert zu überprüfen. Im Vergleich dazu sind die DESI-Tests überraschend traditionell und konservativ.

DESI folgt einer klassischen Aufteilung in Fertigkeiten. Es gibt anders als etwa beim neuen TOEFL keinen Versuch, Fertigkeiten auch kombiniert zu testen. Auch die interkulturellen Kompetenzen werden separat erfasst. Kulturelle und soziokulturelle Aspekte bleiben in den Aufgabenstellungen zum Leseverstehen und zum Hörverstehen ausgeklammert. Der Verzicht auf Fertigkeitskombinationen und die Ausklammerung der interkulturellen Aspekte ist sowohl vom Ziel, Kompetenzmodelle für das Lese- und Hörverstehen zu entwickeln, als auch aus methodischen Gründen verständlich

und vertretbar. Aber trifft das auch für die anderen „konservativen“ Entscheidungen zu?

Die Testhefte zum Leseverstehen enthalten jeweils nur zwei Texte mit Aufgaben. Alle Lesetexte gehören zum Typ der narrativen Texte, und es gibt nur ein einziges Aufgabenformat: Multiple-Choice-Aufgaben. Beim Hörverstehen ist die Auswahl beschränkt auf Audiotexte; Videoaufzeichnungen oder Filme werden nicht verwendet (vgl. Buck 2001). Die Hörtexte liegen, wie die Autoren selbst angeben, am „literate end“ des Kontinuums von Schriftlichkeit/Mündlichkeit (Nold/Rossa in diesem Band). Die Hörtexte werden alle zweimal abgespielt. Einziges Aufgabenformat sind auch hier Multiple-Choice-Aufgaben.

Die Autoren berufen sich für das Testkonstrukt und die Aufgabenentwicklung auf die relevanten Curricula, auf die einschlägige Forschung und auf den Referenzrahmen. Es ist daher gerechtfertigt, die im Vergleich zu anderen neueren Tests sehr traditionelle Ausrichtung der DESI-Tests von diesen drei Bezügen her zu beurteilen. Ich gehe zunächst etwas ausführlicher auf die Tests zum Leseverstehen ein.

Leseverstehen: Textsorten, Leseabsichten und Lesearten

Das DESI-Modul zum Leseverstehen berücksichtigt die Textsorten Erzählung, berichtender Sachtext, Brief und Dramenausschnitt, die alle zum Typ der narrativen Texte gezählt werden. Dies ist, wie die Untersuchungen von Nold/Willenberg (in diesem Band) zeigen, nur ein kleiner Ausschnitt aus dem Spektrum von Texttypen und Textsorten, das die Curricula der Länder vorsehen. Natürlich können Tests immer nur einen Ausschnitt erfassen. Aber man muss sich bewusst machen, was bei dieser Auswahl alles unberücksichtigt bleibt. Es fehlen völlig alle Arten von nichtkontinuierlichen Texten wie Schilder, Listen, Anzeigen, Tabellen oder Grafiken. Solche Textsorten finden sich (wie in der Realität) auch recht häufig in Lehrwerken. Im PISA-Aufgabenpool betrug der Anteil von Leseaufgaben in Verbindung mit nichtkontinuierlichen Texten immerhin fast 40% (Artelt u.a. 2001). Von den übrigen knapp über 60% der Leseaufgaben zu kontinuierlichen Texten entfielen nur rund 12% auf narrative Texte, den einzigen Texttyp bei DESI, wo Textsorten der Typen Darlegung, Beschreibung, Argumentation und Anweisung völlig fehlen.

In der Forschung zum Leseverstehen in der Fremdsprache gibt es einige Hinweise darauf, dass Texte je nach Zugehörigkeit zu einem Diskurstyp eher schwieriger oder leichter zu lesen sind. So gelten expositorische Texte tendenziell als schwieriger als narrative Texte (Alderson 2000). Wichtiger aber ist der in der Literatur hergestellte Zusammenhang zwischen Textsorte und Texttyp auf der einen Seite und Leseabsichten sowie Lesearten auf der anderen Seite (Urquhart/Weir 1998; Weir 2005). Wir lesen z.B. Gebrauchsanweisungen nicht in den gleichen Situationen, nicht mit den gleichen Intentionen und nicht auf die gleiche Art wie Erzählungen oder Dramentexte. Eine Leseart wie das suchende selektive Lesen ist für narrative Texte eher untypisch. Weir schreibt zum Zusammenhang von Texttyp und Lesearten:

> „The relationship between text type and operations being assessed is important. In reading tests for example, if scanning is the focus then collection of description texts containing lots of factual detail are likely to be more suitable than argumentative. Conversely if main ideas are the focus then argumentative texts are likely to contain more macro-propositions than texts full of specific details, i.e., descriptive." (Weir 2005, S. 69)

Alderson, auf den sich die Autoren des Leseverstehenstests für ihr Testkonstrukt beziehen, stellt zwar fest, dass in der Forschung bisher nicht eindeutig nachgewiesen werden konnte, ob Leseziele und Lesearten einen wesentlichen Einfluss haben, rät aber: „Despite the lack of firm evidence of a substantial effect of varying readers' purposes, I would argue that test developers need to consider carefully the tasks they set readers, or the purposes with which their test-takers read (…)." (Alderson 2000, S. 52)

Für Alderson ist die Nähe zu Leseabsichten in realen Lebenssituationen ein wichtiges Kriterium der Validität (ebd.).

Wenn Nold/Rossa (in diesem Band) darauf hinweisen, es sei für das Testkonstrukt von Bedeutung, dass alle Schülerinnen und Schüler zwei Texte (zwei Textsorten) mit Aufgaben bearbeiten und dadurch die „unverzichtbare Vielfalt der Texte" gewährleistet werde, so klingt das recht euphemistisch angesichts des schmalen Ausschnitts aus dem Spektrum der Texttypen und Textsorten. Wenn es dann weiter heißt, dass „im Testkonstrukt auf Grund zeitlicher Beschränkungen unterschiedliche Lesearten nicht systematisch erfasst werden" konnten (ebd.), dann stellt sich die Frage, welche Kriterien für die Auswahl von Lesearten bestimmend waren. Hat die Entscheidung für die Beschränkung auf narrative Texte die Auswahl der berücksichtigten Lesearten bestimmt? Oder wurden bestimmte Lesearten als besonders relevant angesehen und daher Textsorten des narrativen Typs gewählt? Nold/Willenberg (in diesem Band) relativieren die „unverzichtbare Vielfalt der Texte" des Tests im Verhältnis zu derjenigen der curricularen Vorgaben: „Im DESI-Leseverstehensmodul wird allerdings hinsichtlich der Verstehensprozesse (Textdetails oder Hauptaussagen erkennen, erschließen, interpretieren) systematischer und deutlicher differenziert, während die Breite der Textsorten eingeschränkter ist."

Würde man bei der Testentwicklung vom Referenzrahmen ausgehen, dann wäre am Anfang des Auswahlprozesses die Frage zu beantworten, in welchen Situationen bzw. Lebensdomänen Texte gelesen werden können sollen (Europarat 2001). Die Benutzer des Referenzrahmens werden dann aufgefordert zu berücksichtigen und transparent zu machen, „mit welchen Absichten Lernende lesen werden müssen" und „auf welche Art und Weise Lernende lesen werden müssen" (ebd., S. 76).

Während PISA den Ansatz des Referenzrahmens übernimmt und systematisch von den vier Lebensdomänen als Variablen ausgeht, beschränkt sich DESI auf Lesen im privaten Lebensbereich. Der Verzicht auf Lesesituationen des Berufslebens ist bezogen auf die Zielgruppe plausibel. Dagegen ist der Ausschluss von Texten und Lesezwecken des öffentlichen Lebens und des Bildungsbereichs (Lesen, um zu lernen) in

Anbetracht der Präsenz des Englischen im öffentlichen Raum und seiner Rolle für Information und Wissenstransfer über Medien schwer nachzuvollziehen.

Auch wenn, wie bei DESI, das Hauptinteresse auf Verstehensprozesse gerichtet ist, sollte ein möglichst breites Spektrum relevanter Texte als Basis dienen. Alderson konstatiert in seinem Standardwerk „Assessing reading“: „Good tests of reading and good assessment procedures in general will ensure that readers have been assessed on their ability to understand a variety of texts in a range of different topics.“ (Alderson 2000, S. 83).

Leseverstehen: Aufgabenformat

Im DESI-Leseverstehenstest kommt nur ein einziges Aufgabenformat vor, nämlich die Multiple-Choice-Aufgabe. Diese Beschränkung steht nicht nur im Widerspruch zur Vielfalt der Aufgabenformen, die in Lehrplänen und Lehrwerken vorkommen und daher den Lernenden vertraut sind, sondern auch zum Standard, der in der Testliteratur gefordert wird. Alderson formuliert dies sehr deutlich:

> „It is now generally accepted that it is inadequate to measure the understanding of text by only one method, and that objective methods can usefully be supplemented by more subjectively evaluated techniques. Good reading tests are likely to employ a number of different techniques, possibly even on the same text, but certainly across the range of texts tested.“ (Alderson 2000, S. 206)

Alderson konstatiert eine deutliche Tendenz, MC-Aufgaben wenn irgend möglich zu vermeiden (ebd.). Er ist mit der Skepsis keineswegs allein. In der PISA-Untersuchung wurde nur ein geringer Prozentsatz Multiple-Choice-Aufgaben eingesetzt. Weir (2005) referiert oft vorgebrachte Kritikpunkte wie die Schwierigkeit, gute Items zu schreiben, die Ratewahrscheinlichkeit und den problematischen Backwash-Effekt. Er weist auf eine wesentliche Schwäche hin, die gerade für eine Untersuchung bedeutsam ist, die auf die Erfassung von Verstehensprozessen und schwierigkeitsbestimmenden Merkmalen zielt: Aus den Resultaten der MC-Aufgaben ist nicht ersichtlich, ob die richtige Antwort gefunden wurde, indem falsche Antworten ausgeschlossen wurden oder indem die richtige Antwort gewählt wurde. Das bedeutet, dass wir nicht wissen können, ob die richtige Lösung auch ohne die Hilfe durch falsche Distraktoren gefunden worden wäre. Das wichtigste Argument gegen dieses Aufgabenformat ist jedoch die fragliche Testvalidität.

> „It would seem likely that the cognitive processing involved in determining an answer in this format bears little resemblance to the way we process texts for Information in real life, and to the extent that this is the case, they may be considered deficient in terms of theory-based validity.“ (Weir 2005, S. 63)

Wie sehr Multiple-Choice-Aufgaben ein normales Lesen verunmöglichen können, lässt sich leicht in einem Selbstversuch nachvollziehen. In krasser Form wird dies deutlich, wenn man die Aufgaben zum Text „Alicia Keys“ aus dem DESI-Aufgaben-

pool zu lösen versucht. Meine persönliche Erfahrung, die mir von einem Kollegen bestätigt wurde, war, dass man durch die Reihenfolge und die Art der MC-Aufgaben nicht nur zu mehrfachem Lesen, sondern zu einem merkwürdigen Hin- und Her-Lesen gezwungen wird. Das erste Item setzt die Lektüre des ganzen Textes voraus. Die folgenden Items beziehen sich dann mehr oder weniger auf einzelne Abschnitte in der Reihenfolge des Textes, bis dann mit Frage 8 wieder der ganze Text gelesen werden muss, worauf die folgenden Fragen den Leser wieder in die ersten Abschnitte des Textes schicken, um schließlich zu den Schlussabschnitten zu springen. Man ist zwar sicher, den relativ einfachen Text verstanden zu haben, fühlt sich durch die Distraktoren aber verunsichert und zu einem kontrollierenden Vor- und Rückwärts-Lesen genötigt. Man wird zu einer speziellen Art von selektivem Lesen gedrängt, dem nicht eine echte Leseabsicht oder -aufgabe zu Grunde liegt (Suche nach dieser oder jener wichtigen Information), sondern die Suchaufgabe wird durch das Testformat erzeugt: Suche nach der Stelle im Text, welche den Distraktor bestätigt oder falsifiziert. Es ist nicht nur sehr mühsam, sondern auch äußerst zeitaufwändig, die Aufgaben zu lösen.

Abgesehen von der Qualität der einzelnen Multiple-Choice-Aufgaben und ihrer Distraktoren, ist allein schon die Quantität ein Problem. Ich ziehe als Beispiel wieder den Testteil „Alicia Keys“ heran. Der Lesetext umfasst 386 Wörter (1666 Zeichen ohne Leerschläge). Der Text der Multiple-Choice-Aufgaben umfasst 326 Wörter (1569 Zeichen). Durch die Aufgaben werden also rund 90% der Textmenge hinzugefügt. Bei einem zweiten Test aus dem Pool, dem Dramentext, ist die durch die Aufgaben hinzugefügte Textmenge nicht ganz so hoch, liegt aber immerhin noch bei über 70%.

Spolsky (1994), der die verschiedenen Textebenen beschreibt, die in einem Lesetest u.a. durch Anweisungen und Aufgaben zum eigentlichen Lesetest hinzugefügt werden, insistiert darauf, dass der durch Aufgaben hinzugefügte zweite Text von einem anderen Autor, einem anderen Sprecher, stammt. Hinzuzufügen ist, dass der umfangreiche Text der Multiple-Choice-Aufgaben nicht in gleicher Weise ein authentischer Text ist wie der eigentliche Lesetext.

Wie die Aufgabenformate internationaler Diplome und auch die Forschung zeigen, gibt es durchaus Alternativen zu Multiple-Choice-Aufgaben. Andere Aufgabenformen sind vor allem unter dem Aspekt der Validität vorzuziehen[1]:

> „The superiority of the short-answer and Information transfer techniques over all others is that texts can be selected to match performance conditions and test operations appropriate to any level of student, and the techniques are likely to activate almost all the processing elements we discussed earlier in our model

1 Wie der Text von Schröder/Nold zur Rahmenkonzeption von 2002 zeigt, waren die Autoren der DESI-Tests sich durchaus der Problematik von MC-Aufgaben und der höheren Validität anderer Aufgabenformate bewusst. Warum dennoch nicht wie ursprünglich intendiert auch andere Aufgabenformate verwendet wurden, geht aus den mir bekannten Texten nicht hervor.

> of reading. They are accordingly likely to generate the clearest evidence of context- and theory-based validity.“ (Weir 2005, S. 131)

Ähnlich fordert Alderson den Verzicht auf Multiple-Choice-Aufgaben zu Gunsten anderer Aufgabenformate:

> „The challenge for the person constructing reading tests is how to vary the reader’s purpose by creating test methods that might be more realistic than cloze tests and multiple-choice techniques. Admittedly, short-answer questions come closer to the real world (…).“ (Alderson 2000, S. 249)

Nold/Rossa (in diesem Band) erklären mit Hinweis auf den knappen Zeitrahmen, auf die „wünschbare Verwendung von unterschiedlichen Aufgabenformaten“ werde „zugunsten der Testökonomie“ verzichtet. Sicher wäre mehr Zeit für einen Leseverstehenstest wünschenswert. Aber auch in dem begrenzten Zeitrahmen wäre eine Verbindung von Textvielfalt und Vielfalt von Aufgabenformaten zu erreichen. Zum einen könnte Zeit gewonnen werden, indem die extrem zeitraubenden Multiple-Choice-Aufgaben vermieden würden. Zum anderen könnte bei einer so großen Stichprobe wie bei DESI ein Untersuchungsdesign mit mehr durch gemeinsame Aufgaben verankerten Testheften gewählt werden.

Leseverstehen und Textrekonstruktion

Anfänglich wurde im DESI-Projekt der Teil „Textrekonstruktion“ explizit auch als ein Test des Leseverstehens betrachtet (Schröder/Harsch 2002). Dies wurde zwar inzwischen zu Recht revidiert, denn C-Test-Spezialisten stellen in Frage, dass Ergebnisse aus C-Tests nicht nur als Maß allgemeiner Sprachkompetenz, sondern auch als Maß der Lesekompetenz interpretiert werden könnten (z.B. Grotjahn/Klein-Braley/Raatz 2002; Grotjahn 1987, 2002). Aber weiterhin werden auf der Basis der DESI-C-Tests Aussagen zur Fähigkeit der Textrezeption und zu Lesestrategien gemacht (Schröder/Harsch in diesem Band). Auf die Problematik, den Lücken im C-Test bestimmte Prozesse oder Strategien zuzuschreiben, werde ich später noch eingehen.

Hörverstehen

Auch in Bezug auf die Hörverstehenstests lässt sich ähnlich wie beim Leseverstehen feststellen, dass sich die Autoren zwar einerseits bei der Definition des Konstrukts auf neuere Forschung berufen – für das Leseverstehen besonders Alderson (2000), für das Hörverstehen Buck (2001) –, aber dass sie die dort postulierten Konsequenzen für die praktische Testkonstruktion nicht nachvollziehen.

Buck sucht mit seinem Konzept eines „default listening construct“ einen Kompromiss zwischen einem Kompetenz-basierten und einem Task-basierten Ansatz. Mit dem Task-orientierten Ansatz kommt neben den zugrundeliegenden Sprachkompetenzen auch die Rolle des Anwendungskontexts ins Spiel. Nold/Rossa

(in diesem Band) rechtfertigen mit Bezug auf dieses Konzept meiner Meinung nach zu Unrecht die konventionelle Ausrichtung der DESI-Hörverstehenstests, indem sie argumentieren, DESI müsse als Sprachstandserhebung und als Evaluationsstudie die „Situationen der unterrichtlichen Sprachverwendung" im Testkonstrukt als bestimmendes Moment berücksichtigen. Da in den Lehrplänen und im Englischunterricht das Verstehen von „Texten" (wohl verstanden als Gegensatz zu Hörverstehen in Handlungszusammenhängen) im Vordergrund stehe und vor allem didaktisierte, schriftnahe Hörtexte eingesetzt würden, müsse auch der DESI-Test sich an diesem Kontext orientieren. Auch wenn es so sein sollte, dass im Unterricht vor allem nichtauthentische, didaktisierte, schriftnahe Texte für die Förderung des Hörverstehens eingesetzt werden, besagt das noch nicht, dass der Erfolg der didaktischen Bemühungen ebenfalls mit Instrumenten überprüft werden sollte, welche die gleichen Merkmale haben wie die Übungsinstrumente. Zumindest müssten auch reale Anwendungssituationen nach bzw. außerhalb der Schule die Auswahl der Hörtexte und der Aufgaben mitbestimmen.

Dass die Texte der DESI-Hörverstehenstests eher zum „literate end" des „oral/literate continuum" gehören (Nold/Rossa in diesem Band), ist keine notwendige Konsequenz aus dem theoretischen Konstrukt in der Kompromissform, die Buck propagiert. Buck selbst postuliert, dass in Hörverstehenstests gerade solche Texte verwendet werden sollten, welche die typischen Merkmale gesprochener Sprache haben (Buck 2001). Er definiert explizit sein „default listening construct" als die Fähigkeit „to process extended samples of realistic spoken language, automatically and in real time" (ebd. S. 114).

In Bezug auf das Aufgabenformat betont Buck, dass Multiple-Choice-Aufgaben schwierig zu schreiben sind und dass es deutliche Hinweise darauf gibt, dass sie einen starken Methodeneffekt haben. Er empfiehlt, bei einer Testpopulation mit gleicher Ausgangssprache, die Aufgaben in der Erstsprache zu geben. Dies garantiere wahrscheinlich am besten, dass die Resultate für das Hörverstehen nicht durch andere Fertigkeiten „kontaminiert" würden (Buck 2001, S. 143).

Viele internationale Sprachprüfungen sehen anders als die DESI-Tests im Hörverstehensteil nur ein einmaliges Hören vor, andere, z.B. die Cambridge Prüfungen, ein zweimaliges Hören (Brindley/Slatyer 2002). Zweimaliges Hören wird vielfach gefordert, um den Stress in der Testsituation zu mildern. Buck macht darauf aufmerksam, dass zweimaliges Hören die Höraufgabe erleichtert, aber auch wesentlich das Testkonstrukt verändern kann (Buck 2001; vgl. auch Brindley/Slatyer 2002). Es fragt sich, ob bei generellem zweimaligem Hören noch wirklich der Anspruch erhoben werden kann, es werde erfasst, wie der Hörtext „in Echtzeit" verarbeitet wird (Nold/Rossa in diesem Band).

Zwar gibt es widersprüchliche Forschungsergebnisse zur Frage, ob die Schwierigkeit der Aufgabe und das Testresultat dadurch beeinflusst werden (Buck 2001; Brindley/Slatyer 2002), dass die Fragen vor oder nach dem Hören gegeben werden. Aber das seltsame für die DESI-Tests gewählte Verfahren, den Hörtext zunächst einmal ohne Steuerung der Hörintention durch Fragen hören zu lassen und dann vor dem

zweiten Hören die Fragen zu geben, ist kaum eine Lösung des Problems. Denn aus den Resultaten ist nicht ersichtlich, ob das Hören ohne oder mit Frage oder ob einfach das erste oder das zweite Hören die Lösung der Aufgabe ermöglicht hat.

Bei einmaligem Abspielen der Hörtexte könnte Zeit für den Einsatz einer größeren Textvielfalt gewonnen werden. Von realen Anwendungssituationen als auch vom Referenzrahmen her, der für das Hörverstehenskonstrukt herangezogen wird, wäre zu empfehlen, die dialogischen und monologischen Hörtexte zu ergänzen durch Hörtextsorten wie öffentliche Durchsagen oder Anweisungen.

Bei den verwendeten Kurzdialogen wird das Verstehen aus der Lauscherposition, nicht aber das Hörverstehen in der Interaktion erfasst. Inwieweit dies in Zusammenhang mit dem mündlichen Test valide geschieht, lässt sich auf der Basis der vorliegenden Informationen nicht sagen.

Schwierigkeitsbestimmende Merkmale

Die Beschreibung der Merkmale

Die DESI-Untersuchungen sind (nicht nur) für die Forschung besonders interessant, weil für Lese- und Hörverstehen bei der Beschreibung der Aufgabenmerkmale und bei der Ermittlung ihres Einflusses auf die Schwierigkeit gleich vorgegangen wurde. Das erleichtert Vergleiche und die Betrachtung von Gemeinsamkeiten und Unterschieden. Da wo die Formulierung von Aufgabenmerkmalen unterschiedlich ist, fragt sich, ob es sich nur um eine bloße Variation in der Darstellung handelt oder ob es um echte, z.B. fertigkeitsspezifische Unterschiede geht. Dazu einige Beispiele.

Beispiel 1: Beim Leseverstehen sind in den Ausprägungen für den *Textlevel* die Niveaus des Gemeinsamen europäischen Referenzrahmens (Europarat 2001) genannt (A2, B1 usw.). Diese Zuordnung und die Redeweise von der „Kompetenzstufe des Textes“ (Leseverstehen) bzw. der „Schwierigkeitsniveaus der Dialoge“ sind problematisch. Im Handbuch des Europarats für die Zuordnung von Prüfungen zu den Referenzniveaus heißt es deutlich:

> „A text does not have a ‚level‘. It is the competence of the test takers as demonstrated by their responses to the items that can be related to a CEF level. The most that can be said about a text is that it is suitable for inclusion in a test aimed at a particular level.“ (Council of Europe 2003, S. 84)

Beispiel 2: Die sprachlichen Merkmale des Textes sind unter der Bezeichnung Textlevel für Leseverstehen und Textniveau für Textrekonstruktion beschrieben. Man würde für die Abstufungen die gleiche Füllung erwarten. Aber es gibt abgesehen von Formulierungsunterschieden auch Unterschiede, die wohl Einfluss auf die Kodierung haben dürften. Idiomatik ist beispielsweise nur für Texte im Teil Textrekonstruktion ein Kriterium für die Ausprägung fortgeschritten. Während für diese

höchste Schwierigkeitsstufe bei Textrekonstruktion das Kriterium „alle Arten von Satzverknüpfungen“ mit bestimmend ist, ist es beim Leseverstehen das Kriterium „Text mit weniger textverknüpfenden Elementen“. Beim Hörverstehen ist das Merkmal Textlevel aus guten Gründen weggelassen worden. Allerdings gingen damit auch die beim Leseverstehen unter Textlevel berücksichtigten textpragmatischen Aspekte ganz verloren, die es ja auch auf der für das Hörverstehen berücksichtigten Ebene der Textpassage gibt. Es fehlen in der Merkmalsbeschreibung Charakteristika der Mündlichkeit wie Sprecher- und Hörersignale, Gliederungssignale, Wiederholungen, Sprecherwechsel, Simultansprechen usw.

Beispiel 3: Im Vergleich zu Hör- und Leseverstehen sind ausgerechnet in den Merkmalsbeschreibungen für Textrekonstruktion durch C-Tests viel ausführlicher und umfassender differenzierende Ausprägungen von Verstehensprozessen und Strategien aufgeführt. Neuere empirische Forschungen (z.B. Sigott 2004) lassen mehr Vorsicht angeraten sein in Bezug auf die Interpretation der Anwendung von Strategien, Nutzung des Kontexts, lower-level und higher-level processing beim Lösen von C-Tests. Die Lernenden verhalten sich danach beim Füllen der C-Test-Lücken sehr viel unterschiedlicher als in diesen Merkmalsbeschreibungen angenommen wird. Die „Merkmalsbeschreibungen“ sind Hypothesen über mögliche, aber nicht unbedingt notwendige oder wahrscheinliche Prozesse und Strategien. Da nicht klar ist, welches Verhalten die Lücken evozieren, lässt sich auch nicht auf eine latente Disposition schließen. Außerdem sind die Formulierungen teilweise sehr vage und problematisch, z.B. „bestimmte Abschnitte verlangen mentale Modelle“, „mentale Modellbildung zum Verständnis notwendig“ oder „mittlerer Bedarf an Weltwissen und Interpolation nötig“ oder gar „Lücken, zu deren Schließung es komplexer Informationsverarbeitung bedarf“ (Harsch/Schröder in diesem Band).

Eine präzise und griffige Beschreibung der Aufgabenmerkmale ist eine wichtige Voraussetzung für eine verlässliche Zuschreibung der Ausprägungsstufen zu den Items und damit für die Kodierung[2]. Die zweite Voraussetzung ist ein intensives Training derjenigen, die diese Einschätzung vornehmen. Das zeigt auch die Erfahrung im so genannten „Grid-Projekt“ (Alderson u.a. 2004). Ausgehend von der Kritik, dass die holistischen Könnens-beschreibungen des Referenzrahmens für Testautoren zu wenig präzise seien, wurde in diesem Projekt ein Raster zur analytischen Beschreibung von Texten und Aufgaben für die rezeptiven Fertigkeiten entwickelt. Bei der Anwendung des Rasters zeigte sich dann, dass selbst Testexperten auch bei der Verwendung eines solchen analytischen Instrumentariums (oder gerade wegen des analytischen und nicht holistischen Zugriffs) sehr unterschiedlich einschätzen, was bestimmte Aufgaben prüfen und welchem Niveau sie zuzuordnen sind. Daher wird ein intensives Rater-Training postuliert.

2 Die Formulierung der Aufgabenmerkmale und ihrer Ausprägungen hat sich offenbar im Verlauf des DESI-Projekts weiterentwickelt. Mir ist nicht recht klar, ob die in diesem Band enthaltenen schlanken Formulierungen für die Merkmalsausprägungen Zusammenfassungen darstellen und mit Anweisungen für die Kodierung identisch sind oder ob für die Kodierung andere, differenziertere Deskriptoren verwendet wurden.

Man darf deshalb gespannt sein auf die Ergebnisse der angekündigten Validierungsstudien, in denen untersucht wird, ob auch andere Personen als die Autoren, welche sowohl die Testaufgaben als auch die Aufgabenmerkmale formuliert haben, die Aufgaben in vergleichbarer Weise interpretieren (kodieren) können (Nold/Rossa; Harsch/Schröder in diesem Band).

Die schwierigkeitsbestimmenden Faktoren

Die zahlreichen Untersuchungen zur Ermittlung von Faktoren, welche die Schwierigkeit von Aufgaben beeinflussen, haben teilweise unterschiedliche Faktoren und Faktorenbündel identifiziert (Bygate/Skehan/Swain 2001; Bachman 2002; Brindley/Slatyer 2002). Die Ergebnisse der DESI-Untersuchungen, welche Merkmale wesentlichen Einfluss auf die empirischen Schwierigkeitswerte haben und die daher bei der Bestimmung der Schwellenwerte für die Kompetenzniveaus berücksichtigt wurden, bestätigen teilweise frühere Untersuchungen, sind teilweise aber auch unerwartet. So erstaunt es, dass Hörverstehen und Leseverstehen nur zwei – zweifellos wichtige – schwierigkeitsbestimmende Merkmale gemeinsam haben: M1 – den inhaltlichen Fokus der Aufgabe (konkret vs. abstrakt) und M3 – die Verstehensabsichten.

Ohne die Kenntnis aller Aufgaben, der vorgenommenen Kodierungen, der Daten und aller je Niveau vorkommenden Merkmalskombinationen lassen sich die Resultate schwer beurteilen. Aber man darf vermuten, dass die Ergebnisse in Bezug auf die Dimensionen und Stufen der Skalen zu einem gewissen Teil auch dadurch zu erklären sind, dass der Aufgabenpool für Lese und Hörverstehen relativ klein war und wenig Variation in Bezug auf Text- und Aufgabentyp wie auf Text- und Aufgabenschwierigkeit enthielt.

Skalen mit Kompetenzbeschreibungen

Die für das Leseverstehen und Hörverstehen entwickelten vier Kompetenzniveaus erscheinen nicht zuletzt deshalb attraktiv und interessant, weil die Niveauabgrenzung erstens auf einer Kombination von Merkmalsausprägungen beruht und zweitens weil es sich um eine überschaubare Zahl von relevanten schwierigkeitsbestimmenden Merkmalen handelt. Allerdings wird noch viel zusätzliche Erklärungs- und Interpretationsarbeit nötig sein, um die Kompetenzniveaus (Tabelle 2 der Beiträge von Nold/Rossa in diesem Band) auch für die verschiedenen Interessentengruppen wie Bildungspolitiker, Lehrerinnen und Lehrer, Curriculumentwickler, Lehrwerkautoren oder gar die Schülerinnen und Schüler verständlich zu machen. Fragen, die Leser stellen werden, sind z.B.: Wie ist es zu interpretieren, dass die (in Logits ausgedrückten) Abstände zwischen den Niveauschwellen so unterschiedlich groß sind – innerhalb der Skala für Lese- bzw. Hörverstehen und unterschiedlich im Vergleich der zwei Skalen? Wie ist es zu interpretieren, dass abgesehen vom untersten Niveau *KN A* (viermal die Ausprägung 0) und vom obersten Rand des Niveaus *KN D* (vier-

mal die Ausprägung 2) die Kombinationen der Merkmalsausprägungen beim Lese- und Hörverstehen doch recht unterschiedlich sind? Inwieweit handelt es sich um fertigkeitsspezifische Unterschiede oder um Effekte des Testdesigns? Eine Hilfe für das Verständnis der Kompetenzniveaus sind die Kann-Beschreibungen.

Ein Ziel des DESI-Projekts war es, zu anschaulichen Beschreibungen von Kompetenzniveaus zu kommen. Die Beiträge zu den rezeptiven Fertigkeiten münden denn auch in Kann-Beschreibungen zu den ermittelten Kompetenzniveaus.

Kann-Beschreibungen sind durch den Gemeinsamen europäischen Referenzrahmen (Europarat 2001) und durch das Europäische Sprachenportfolio vertraut geworden. Dass diese Kompetenzbeschreibungen offensichtlich für eine sehr große Zahl von Lernenden und Lehrenden plausibel sind und von ihnen sinnvoll verwendet werden können, liegt wahrscheinlich an einigen charakteristischen Eigenschaften der Deskriptoren des Referenzrahmens. Als „gut" und gut skalierbar erwiesen sich nach Schneider/North (2000, S. 89) Kompetenzbeschreibungen, die folgende Bedingungen erfüllen: Das Können ist positiv formuliert; sie machen für sich allein genommen Sinn; ihre Interpretation ist nicht abhängig von anderen Beschreibungen des gleichen Niveaus oder von Beschreibungen angrenzender Niveaus; Niveauunterschiede sind nicht nur ausgedrückt durch verbale Abstufungen; sie enthalten wenig Jargon und Fachterminologie; sie sind konkret, klar und kurz.

Diese Eigenschaften machen die Kompetenzbeschreibungen für die Betroffenen, d.h. auch für Laien, verständlich und attraktiv. Einzelne der positiven Merkmale erscheinen jedoch möglicherweise für bestimmte Benutzergruppen, besonders Fachleute wie Testspezialisten, als Nachteil. In kurzen Deskriptoren beispielsweise können nicht viele der Faktoren aufgeführt werden, die einen Handlungsbereich und ein Niveau mitbestimmen. So enthalten z.B. die einzelnen Deskriptoren des Referenzrahmens zum Leseverstehen nicht jeweils systematisch Angaben zu Leseintention, Lesesituation, Textsorte, Diskurstyp, Textlänge, Themenbereich, Informationsdichte, sprachlichen Merkmalen der Texte usw., sondern sie fokussieren Schlüsselmerkmale und nennen nur das, was für die sprachliche Handlungskompetenz auf dem entsprechenden Niveau ganz besonders typisch ist.

Testautoren, die Prüfungsaufgaben zu den Deskriptoren des Referenzrahmens entwickeln, möchten jedoch die Faktoren kontrollieren, welche die Schwierigkeit einer Aufgabe beeinflussen. Für sie stellt sich die Aufgabe, die Deskriptoren zu interpretieren, für ihre Zwecke anzupassen und als Anweisungen für die Ausarbeitung von Test-Items systematisierend zu erweitern (Alderson u.a. 2004).

Der Referenzrahmen und entsprechend das Sprachenportfolio enthalten sowohl aufgabenorientierte als auch beurteilungs- bzw. diagnoseorientierte Deskriptoren. In vielen Deskriptoren sind die Aufgabenbeschreibungen mit qualitativen Aussagen verbunden. Denn Checklisten und Skalen für die kontinuierliche Beurteilung durch Lehrende oder für die Selbstbeurteilung funktionieren am besten, wenn sie nicht nur aussagen, was Lernende tun können (aufgabenorientiert), sondern auch, wie gut sie es können (Europarat 2001). Rein aufgabenorientierte Skalen sind in erster Linie für

Testautoren bestimmt. Sie beschreiben Aufgaben, welche die Lernenden lösen können sollen. Der Akzent liegt darauf, was die Lernenden tun können.

Während die Kompetenzbeschreibungen des Referenzrahmens empirisch kalibriert wurden, beruhen Erweiterungen und Anpassungen von Skalen des Referenzrahmens in der Regel nur auf Expertenurteilen ohne empirische Validierung. Bei vorsichtigen Bearbeitungen wurden möglichst viele Elemente aus kalibrierten Deskriptoren des jeweiligen Niveaus unverändert übernommen und neu kombiniert (Schneider/Lenz 2001; Lenz/Schneider 2004). Auch die neuen Raster und Skalen im „Manual" für die Zuordnung von Sprachprüfungen zu den Referenzniveaus (Council of Europe 2003) wurden alle durch eine Kompilation von skalierten Deskriptoren gewonnen, die in verschiedenen Einzelskalen des Referenzrahmens enthalten sind.

Für die Zuordnung von Testaufgaben zu den Referenzniveaus sieht das „Manual" ein anspruchsvolles Verfahren vor. Die an der Zuordnung Beteiligten sollten sich durch verschiedene Aktivitäten intensiv mit den Niveaubeschreibungen vertraut machen, die eigenen Tests differenziert in Bezug auf Kategorien des Referenzrahmens beschreiben und sich in einem Verfahren der Standardsetzung in der Einschätzung von Musteraufgaben zum Lese- und Hörverstehen eichen, um dann die eigenen Testaufgaben in Beziehung zu den Musteraufgaben zu setzen. Erst wenn diese Musteraufgaben für die Referenzniveaus vorliegen, wird es möglich sein, die eingeschätzten DESI-Aufgaben über einen Vergleich mit den Musteraufgaben zu den Skalen des Referenzrahmens in Beziehung zu setzen.

Schon jetzt kann man der Frage nachgehen, welchen Status die DESI-Skalen haben und in welcher Beziehung sie zum Referenzrahmen stehen. Hier ist kein Raum für eine eingehende Analyse. Einige Hinweise sollen genügen.

Es wird der Anspruch erhoben, dass die Kann-Beschreibungen die Aufgabenmerkmale „auf Situationen der Sprachverwendung außerhalb der Testsituation" hin generalisieren (Nold/Rossa in diesem Band). Ich möchte dazu stichwortartig einige Beobachtungen mit kurzen Kommentaren zusammenstellen:

- Die Kann-Beschreibungen zu Hörverstehen, Leseverstehen und Textrekonstruktion unterscheiden sich in der Ausführlichkeit und teilweise im Stil der Formulierungen; am kürzesten sind diejenigen zum Lesen, ausführlicher, in Punkte gegliedert und teilweise einfacher formuliert die zum Hören und am längsten und stark (fach-)jargonhaft die zur Textrekonstruktion.
- Die Kann-Beschreibungen beschränken sich nicht nur auf diejenigen Aufgabenmerkmale und Ausprägungen der Merkmale, die auf der Regressionsanalyse basierend als wesentliche schwierigkeitsbestimmende Merkmale für die Festlegung der Niveauschwellen dienten. Die Analysen legten nach Nold/Rossa (in diesem Band) nahe, teilweise nur zwei Ausprägungen je Merkmal zu unterscheiden, weshalb entweder die beiden unteren Ausprägungen (Kodierungen 0 und 1) oder die oberen Ausprägungen (Kodierungen 1 und 2) zusammengefasst wurden. In den Kann-Beschreibungen jedoch sind wieder die nicht zusammengelegten Ausprägungen in die Formulierungen der Niveaucharakterisierung eingegangen. Es fragt sich, inwieweit verbale Abgrenzungen

zwischen „abstraktere Einzelinformationen“, „eine begrenzte Anzahl abstrakterer Informationen“ und „abstrakte Informationen“ tatsächlich auf reale Unterschiede verweisen. Oder: Was ist der reale Unterschied zwischen einem Text mit „weniger frequentem Wortschatz“ und einem Text mit „erweitertem Wortschatz“?

- Die Kann-Beschreibungen bleiben unterschiedlich nahe an den Deskriptoren für die Aufgabenmerkmale.
- Während nach den Kann-Beschreibungen für die beiden ersten Niveaus des Leseverstehens das unterscheidende Merkmal darin besteht, dass man auf dem ersten Niveau (*KN A*) „konkrete Einzelinformationen“ verstehen kann, dagegen auf dem zweiten Niveau (*KN B*) „abstraktere Einzelinformationen in alltäglichen Kontexten“, sollen beim Hörverstehen solche „abstraktere Einzelinformationen in alltäglichen Kontexten“ erst ein Charakteristikum für das dritte Niveau (*KN C*) sein. Bei diesen und ähnlichen Unterschieden in den Kann-Beschreibungen für das Lese- und Hörverstehen fragt es sich, ob es sich hier um reale fertigkeitsspezifische Unterschiede oder um Effekte des Testdesigns handelt.
- In die Kann-Beschreibungen zu Textrekonstruktion sind 1. viele Hypothesen zu denkbarem Lösungsverhalten eingegangen, was allerdings noch keinen Rückschluss auf dieses oder jenes Können erlaubt; 2. sind in diese Kann-Beschreibungen viele Elemente aus Formulierungen der Skalen des Referenzrahmens teils wörtlich, teils abgeändert eingebaut worden, allerdings in m.E. problematischer Verallgemeinerung, beispielsweise wenn Formulierungselemente aus Skalen des Referenzrahmens zum Leseverstehen übernommen werden wie „Kann auch komplexere fiktive Texte und Sachtexte rezipieren“ oder „Kann auch komplexe, anspruchsvolle Texte verstehen“ (Harsch/Schröder in diesem Band).

Die vorliegenden Kann-Beschreibungen sind wohl als „Work in Progress“ zu sehen. Es müsste klar kommentiert und begründet werden, welchen Status sie beanspruchen: den einer auf konkrete Tests bezogenen Beschreibung oder den von verallgemeinerten, auf Realsituationen ausgelegten Kompetenzbeschreibungen. Wenn die Kompetenzbeschreibungen nicht nur für den Kreis der Spezialisten bestimmt sein sollen, müssten sie verständlicher und leserfreundlicher umformuliert werden. Die vorliegenden Kompetenzbeschreibungen lassen sich nicht leicht in Kann-Beschreibungen für die Selbstbeurteilung umformulieren. Ich vermute, dass sich viele Lehrerinnen und Lehrer und erst recht viele Schülerinnen und Schüler wenig vorstellen können unter Formulierungen wie „Kann abstrakte Informationen (z.B. Meinungen, Textstrukturen) mit Hilfe von Inferieren impliziter Informationen verknüpfen oder sehr komplexe Einzelinformationen interpretieren (…)“. In einem nächsten Schritt sollte – möglichst mit Anwendung von Verfahren des „Manual“ – aufgezeigt werden, in welcher Beziehung die Kompetenzbeschreibungen zu den Niveaus des Referenzrahmens stehen.

Einige Schlussfolgerungen

Welche Konsequenzen wären für nachfolgende Projekte zu ziehen?

- Die Kompetenzbeschreibungen des Referenzrahmens sollten nicht erst in der Phase der Generalisierung ins Spiel gebracht werden, sondern als Ausgangspunkt für die Planung der Text- und Aufgabenauswahl genutzt werden. Als Hilfe für eine begründete und transparente Testspezifikation können die Checklisten im „Manual" des Europarats dienen (Council of Europe 2003).
- Es wäre zu überlegen, ob die Testaufgaben sich an den klassischen vier Fertigkeiten ausrichten sollten oder ob dem Referenzrahmen folgend Interaktion und damit die Fertigkeitskombination stärker betont werden und auch Aufgaben zur Sprachmittlung mit einbezogen werden sollen.
- Es empfiehlt sich, die Deskriptoren des Referenzrahmens, die auf die Welt der Erwachsenen hin ausgelegt sind, vor der Ausarbeitung der Tests dem Alter der Zielgruppe entsprechend zu adaptieren und die adaptierten Deskriptoren mit den im Referenzrahmen beschriebenen Verfahren zu validieren.
- Um die Validität der Tests zu gewährleisten sollten verschiedene relevante Texttypen und Aufgabenformate vorgesehen werden. Ein breiteres Spektrum und damit ein größerer Aufgabenpool (Klieme u.a. 2003) ist auch bei begrenzt zur Verfügung stehender Testzeit möglich, wenn mehr unterschiedliche, durch gemeinsame Aufgaben verankerte Testhefte eingesetzt werden.
- Es sollte möglichst mehr Zeit eingeplant werden als bei den DESI-Tests. Es kann aber auch Zeit gewonnen werden, wenn beim Leseverstehen nicht so zeitaufwändige Multiple-Choice-Aufgaben verwendet werden und wenn beim Hörverstehen nur einmaliges Hören vorgesehen wird.
- Die Auswahl der Aufgaben soll in Kenntnis der Lehrpläne vorgenommen werden, aber das Spektrum der Textsorten und -aufgaben darf nicht übervorsichtig auf das beschränkt bleiben, was in allen Lehrplänen enthalten ist. Nachuntersuchungen zu PISA haben deutlich gezeigt, dass die Ergebnisse nur sehr geringfügig anders ausgefallen wären, wenn nur lehrplanvalide Aufgaben eingesetzt worden wären (Kunter u.a. 2002; Moser/Berweger 2003). Wenn möglich sollten Relevanz und Akzeptanz der gewählten Texte und Aufgaben mit Fachdidaktikern und Praktikern z.B. in Form von Workshops geklärt werden.
- C-Tests sollten nicht als Messinstrument für Lesestrategien eingesetzt werden. Als Tests allgemeiner Sprachkompetenz, die mit fertigkeitsbezogenen Tests teilweise hoch korrelieren, könnten sie vielmehr zur Verankerung von Testteilen und in der Entwicklungsphase zur Kalibrierung von Aufgaben verwendet werden (vgl. Arras/Grotjahn 2002).
- Die aus den Beschreibungen von Aufgabenmerkmalen und den Testresultaten gewonnenen Könnensbeschreibungen sollten mit einer größeren Gruppe von Praktikern und Experten validiert werden, und zwar nicht allein durch Stellungnahmen, sondern auch z.B. durch Sortieraufgaben oder ähnliche Aktivitäten.

- Die Zuordnung von Testaufgaben zu den Niveaus des Referenzrahmens kann nicht nur Sache der Testautoren sein, sondern sollte, wie es das „Manual“ des Europarats empfiehlt, in einer größeren Gruppe intensiv eingeübt und mit Verfahren der kontrollierten Standardsetzung vorgenommen werden.

Auch wenn die DESI-Tests als Forschungsinstrument gedacht sind und nicht als Instrumente, die in den Unterricht gehören, sollte die Gefahr von Rückwirkungen auf die Einstellung von Verantwortlichen für den Fremdsprachenunterricht, auf Lehrpersonen und auf den Unterricht selbst nicht unterschätzt werden.

Literatur

Alderson, J.C. (2000): Assessing Reading. Cambridge: CUP (= Language Assessment).

Alderson, J.C./Figueras, N./Kuijper, H./Nold, G./Takala, S./ Tardieu, C. (2004): Reading and Listening. Final Report of The Dutch CEF Construct Project. Online: http://www.ealta.eu.org/

Arras, U./Grotjahn, R. (2002): TestDaF: Aktuelle Entwicklungen. In: Fremdsprachen und Hochschule 66, S. 65-88.

Artelt, C./Baumert, J./Klieme, E./Neubrand, M./Prenzel, M./Schiefele, U./Schneider, W./Schümer, G./Stanat, P./Tillmann, K.-J./Weiß, M. (Hrsg.) (2001): PISA 2000. Zusammenfassung zentraler Befunde. Berlin: Max-Planck-Institut für Bildungsforschung. Online: www.pisa.oecd.org.

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

Brindley, G./Slatyer, H. (2002): Exploring task difficulty in ES Listening assessment. In: Language Testing 19, S. 369-394.

Buck, G. (2001): Assessing Listening. Cambridge: University Press.

Bygate, M./Skehan, P./Swain, M. (Hrsg.) (2001): Researching Pedagogic Tasks. Second Language Learning, Teaching and Testing. Harlow: Pearson Education (= Applied Linguistics and Language Study).

Council of Europe (2003): Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF). Manual, Preliminary Pilot Version. Strasbourg: Language Policies. http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/

Europarat (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Berlin u.a.: Langenscheidt. Online-Version: http://www.goethe.de/referenzrahmen

Grotjahn, R. (1987): Ist der C-Test ein Lesetest? In: Addison, A/Vogel, K (Hrsg.): Lehren und Lernen von Fremdsprachen im Studium. Bochum: AKS, S. 230-248.

Grotjahn, R. (2002): Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. In Grotjahn, R. (Hrsg.): Der C-Test. Theoretische Grundlagen und praktische Anwendungen. Band 4. Bochum: AKS (= FLF Fremdsprachen in Lehre und Forschung 32), S. 211-221.

Grotjahn, R./Klein-Brayley, C./Raatz, U. (2002): C-Tests: an overview. In Coleman, J.A./ Grotjahn, R./Raatz, U. (Hrsg): University Language Testing and the C-Test. Bochum: AKS (= FLF Fremdsprachen in Lehre und Forschung 31), S. 93-114.

Klieme, E./Avenarius, H./Blum, W./Döbrich, P./Gruber, H./Prenzel, M./Reiss, K./Riquarts, K./ Rost, J./Tenorth, H.-E./Vollmer, H. J. (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Hrsg. Bundesministerium für Bildung und Forschung. Bonn: BMBF.

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2004): Bildungsstandards für die erste Fremdsprache (Englisch/ Französisch) für den Mittleren Schulabschluss. Neuwied: Luchterhand. Online: http://www.kmk.org/doc/beschl/aschulw.htm

Kunter, M./Schümer, G./Artelt, C./Baumert, J./Klieme, E./Neubrand, M./Prenzel, M./Schiefele, U./Schneider, W./Stanat, P./Tilmann, K.-J./Weiß, M. (Hrsg.) (2002): PISA 2000. Dokumentation der Erhebungsinstrumente. Berlin: Max-Planck-Institut für Bildungsforschung. Materialien aus der Bildungsforschung Nr. 72.

Lenz, P./Schneider, G. (2004): A bank of descriptors for self-assessment in European Language Portfolios. Strasbourg: Council of Europe, Language Policy Division. Online www.coe.int/portfolio

Moser, U./Berweger, S. (2003): Lehrplan und Leistungen. Thematischer Bericht der Erhebung PISA 2000. Hrsg. Bundesamt für Statistik (BFS) und Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). Neuchâtel. Online: www.pisa.admin.ch.

Schneider, G. (2005, im Druck): Der „Gemeinsame europäische Referenzrahmen für Sprachen" als Grundlage von Bildungsstandards für die Fremdsprachen – Methodologische Probleme der Entwicklung und Adaptierung von Kompetenzbeschreibungen. In: Schweizerische Zeitschrift für Bildungswissenschaften 1.

Schneider, G./Lenz, P. (2001): Guide for Developers of a European Language Portfolio. Strasbourg: Council of Europe. Online: www.coe.int/portfolio >Documentation

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

Schröder, K./Harsch, C. (2002): Rahmenkonzeption zur Erfassung sprachlicher Kompetenzen. Teil II: Testkonzeptionen im Englischen. Einordnung in das Rahmenkonzept. Online: http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/didaktik/forschung/desi/

Schröder, K./Nold, G. u.a. (2002): Rahmenkonzeption zur Erfassung sprachlicher Kompetenzen. Teil I: Theoretische Grundlegung zur Konzeptualisierung sprachlicher Kompetenzen und ihrer Operationalisierung im Englischen. Online: www.philhist.uni-augsburg.de/lehrstuehle/anglistik/didaktik/forschung/desi

Sigott, G. (2004): Towards Identifying the C-Test Construct. Bern u.a.: Lang (= Language Testing and Evaluation 1).

Spolsky, B. (1994): Comprehension testing, or can understanding be measured? In: Brown, G./Malmkjaer, K./Pollitt, A./Williams, J. (eds.): Language and understanding. Oxford: University Press, S.141-152.

Urquhart, S.A.H./Weir, C.J. (1998): Reading in a second language: process, product and practice. London/New York: Longman (= Applied Linguistics and Language Study).

Vollmer, H.J. (2003): Ein gemeinsamer europäischer Referenzrahmen für Sprachen: Nicht mehr, nicht weniger. In: Bausch, K.-R./Christ, H./Königs, F.G/Krumm, H.-J. (Hrsg.): Der Gemeinsame europäische Referenzrahmen in der Diskussion. Tübingen: Narr, S. 192-206.

Weir, C.J. (2005): Language Testing and Validation: An Evidence-Based Approach. Basingstoke: Palgrave Macmillan (= Research and Practice in Applied Linguistics)