

TestDaF: Aktuelle Entwicklungen*

Ulrike Arras

Rüdiger Grotjahn

1 Vorbemerkung

Seit 2001 können ausländische Studienbewerberinnen und -bewerber, die an einer Hochschule in Deutschland studieren wollen, den Nachweis über ausreichende Sprachkenntnisse mit dem neu entwickelten „Test Deutsch als Fremdsprache“ (TestDaF) erbringen. Beim TestDaF handelt es sich um einen standardisierten Test, der Sprachkenntnisse im akademischen Kontext im oberen Leistungsspektrum, d.h. auf fortgeschrittenem Niveau, misst. Die Prüfung kann bereits im Heimatland abgelegt werden und erleichtert damit den Zugang zu einem Hochschulstudium in Deutschland. Der TestDaF ist somit ein Instrument zur Förderung des Studien- und Wissenschaftsstandorts Deutschland. Der Test misst die vier Fertigkeiten Lesen, Hören, Schreiben und Sprechen getrennt voneinander in vier Subtests. Der Testteil „Leseverstehen“ besteht aus drei Texten mit insgesamt 30 Items (Zuordnungs-, Mehrfachwahl- und Auswahl-Items), der Testteil „Hörverstehen“ aus 25 Items (gesteuerte Notizen, Alternativformen). Die Überprüfung der schriftlichen Ausdrucksfähigkeit erfolgt anhand einer Texterstellungsaufgabe. Zur Bewertung der mündlichen Ausdrucksfähigkeit wird ein „Simulated Oral Proficiency Interview“ (SOPI) eingesetzt, das aus 10 situativ eingebetteten Aufgaben besteht, in denen die Prüfungsteilnehmenden sprachlich reagieren müssen (vgl. Kniffka & Üstünsöz-Beurer 2001).¹ Die individuelle Leistung wird den drei TestDaF-Niveaustufen (TDN) 3, 4 und 5 zugeordnet. Unterhalb von TDN 3 differenziert der TestDaF nicht; eine Einstufung auf unter TDN 3 besagt daher lediglich, dass das Eingangsniveau von TestDaF nicht erreicht wurde. Auf dem Zeugnis werden die erbrachten Leistungen nach Fertigkeiten getrennt ausgewiesen, so dass die Prüfungsteilnehmenden ebenso wie die Hochschulen und Fachbereiche ein differenziertes Bild über die Stärken und evtl. Schwächen in den verschiedenen Fertigungsbereichen erhalten. Die TestDaF-Leistungstufen orientieren sich am „Gemeinsamen Europäischen Referenzrahmen für Sprachen“ des Europarats

* Erweiterte Fassung eines Vortrags auf der 22. Arbeitstagung des AKS in Chemnitz, 28. Februar 2002 (vgl. Arras/Grotjahn erscheint). Der Abschnitt 4 stellt eine stark gekürzte Fassung von Arras/Eckes/Grotjahn (2002) dar.

¹ Zur Beschreibung des Formats und der Anforderungen getrennt nach den einzelnen Prüfungsteilen siehe <http://www.testdaf.de>.

sowie an der Skala der *Association of Language Testers in Europe* (ALTE). Das TestDaF-Leistungsspektrum liegt nach den Kompetenzskalen des Europarats ungefähr im Bereich B 2.1 (*Lower Vantage Level*) bis C 1.2 (*Higher Effective Proficiency*) und nach den Skalen der ALTE ungefähr im Bereich ALTE-Stufe 3 (*Independent User*) bis 5 (*Good User*). Zudem misst der TestDaF Sprachverwendung im akademischen Kontext, während insbesondere der Referenzrahmen eher allgemeine Sprachfähigkeit erfasst.²

Da es sich beim TestDaF um einen High-Stakes-Test handelt, ist die Einhaltung testwissenschaftlicher Standards unter Einschluss von Fairness von zentraler Bedeutung. Die Prüfung wird zu festgesetzten Terminen weltweit an lizenzierten Testzentren abgelegt. Die Erstellung der Testaufgaben, die Auswertung der Prüfungsdaten sowie die testmethodische Absicherung und die wissenschaftlich gestützte Weiterentwicklung des Testformats erfolgen zentral im TestDaF-Institut.³

2 Einige Daten zur Entwicklung

Das TestDaF-Institut wurde im Februar 2001 gegründet und ist eine Einrichtung der Gesellschaft für Akademische Testentwicklung e.V. Gegründet wurde das TestDaF-Institut von Einrichtungen, die die deutschen Hochschulen vertreten und mit auswärtiger Kulturpolitik befasst sind (die Hochschulrektorenkonferenz, der Deutsche Akademische Austauschdienst e.V. sowie das Goethe-Institut Inter Nationes e.V.), und von Institutionen, die das TestDaF-Institut wissenschaftlich beraten (die FernUniversität – Gesamthochschule in Hagen, die Ruhr-Universität Bochum, die Universität Leipzig und der Fachverband Deutsch als Fremdsprache). Seit der Einführung des TestDaF wurden ca. 220 Testzentren in ca. 70 Ländern lizenziert. Über 2500 Personen haben an drei Prüfungsterminen den TestDaF bisher abgelegt, wobei derzeit die größte Nachfrage in Osteuropa und in Deutschland zu verzeichnen ist. Die relativ große Zahl an Teilnehmenden in Deutschland erklärt sich z.T. aus der Situation, dass der TestDaF im ersten Jahr in vielen Ländern noch nicht durchgeführt werden konnte. Die mit Abstand stärkste Nachfrage wird in Zu-

² Niveaustufen-Beschreibung und *Can-Do-Statements* der ALTE siehe <http://www.alte.org>. Die Niveaustufen-Beschreibungen des Gemeinsamen Europäischen Referenzrahmens siehe <http://www.sprachenportfolio.ch/> und <http://www.goethe.de/z/50/commeuro/> sowie Europarat (2001: Kap. 3). Zur Beschreibung der TestDaF-Niveaustufen siehe <http://www.testdaf.de>.

³ Das TestDaF-Institut ist ein An-Institut an der FernUniversität – Gesamthochschule in Hagen und an der Ruhr-Universität Bochum. Weitere Informationen zum TestDaF geben z.B. Bolton (2000), Grotjahn/Kleppin (2001) und Projektgruppe TestDaF (2000).

kunft in Ostasien zu verzeichnen sein. Für die nähere Zukunft ist daher geplant, vor allem auch in der VR China den TestDaF anzubieten.

Im Folgenden wollen wir einige aktuelle Entwicklungen und Überlegungen vorstellen. Zunächst sollen Probleme bei der Bewertung des Prüfungsteils *Schriftlicher Ausdruck* dargestellt werden. Sodann wird die Möglichkeit des Einsatzes von C-Tests zur Verankerung von TestDaF-Aufgaben diskutiert. Schließlich soll kurz auf die Entwicklung einer computerbasierten Version des TestDaF eingegangen werden.

3 Bewertung des Prüfungsteils *Schriftlicher Ausdruck*

3.1 Zum Aufgabenformat

Der Prüfungsteil *Schriftlicher Ausdruck* (siehe das Beispiel im Anhang) besteht aus einer Textstellungsaufgabe. Für die Bewältigung der Aufgabe stehen 60 Minuten zur Verfügung; eine Vorbereitungszeit oder Hilfsmittel sind nicht vorgesehen. Bei den geforderten Schreibhandlungen handelt es sich um solche, die typisch für das wissenschaftliche Schreiben sind: Beschreiben und Zusammenfassen von Informationen aus einer grafischen Vorlage, Paraphrasieren von Fremdmeinungen, Diskussion unterschiedlicher Meinungen oder unterschiedlicher Überlegungen, begründetes Argumentieren, Darstellung des eigenen Standpunktes.

3.2 Zum Bewertungsverfahren bei den produktiven Fertigkeiten

Die produktiven Fertigkeiten *Schriftlicher Ausdruck* und *Mündlicher Ausdruck* werden von speziell geschulten Korrektorinnen und Korrektoren bewertet. Vor jeder Korrekturphase finden Standardisierungstreffen statt, bei denen die Korrektorinnen und Korrektoren anhand der zu bewertenden Aufgaben und anhand einer Auswahl an mündlichen und schriftlichen Leistungen kalibriert werden. Damit soll erreicht werden, dass bei der konkreten Auswertung der mündlichen und schriftlichen Leistungen möglichst einheitliche Bewertungsmaßstäbe angelegt werden.

Das standardisierte Bewertungsverfahren sieht vor, dass die von den Prüfungsteilnehmenden erstellten Texte anhand vorgegebener Deskriptoren (Bewertungskriterien) den verschiedenen TestDaF-Niveaustufen zugeordnet werden. Hierbei ist zu beachten, dass ein Text zunächst getrennt nach drei übergeordneten Kriterien bewertet wird. Diese Kriterien erfassen zum einen die Gesamtwirkung bei der Rezeption des Textes. Es handelt sich um eine eher holistische Erfassung der Leistung. Zum anderen werden die sprachliche und die inhaltliche Umsetzung der Aufgabe bewertet, was ein eher analytisches Vorgehen erfordert. Im Anschluss ergeben diese drei Einstufungen anhand der drei gleichwertigen Kriterien die Gesamtbewertung, wobei festgelegt ist, dass im Zweifelsfalle in das höhere Niveau einzustufen ist. Erzielt beispielsweise ein Text die Einstufungen „Gesamteindruck: TDN 5“, „Behandlung der Aufgabe: TDN 5“ und „Sprachliche Realisierung: TDN 4“, so ergibt sich eine Gesamtbewertung auf TDN 5.

Alle Texte werden von zwei verschiedenen Korrektorinnen oder Korrektoren unabhängig voneinander bewertet. Im Falle einer uneinheitlichen Einstufung, d.h. wenn die beiden Urteile um eine oder mehrere Stufen auseinander liegen, wird eine Drittkorrektur durchgeführt. Diese endgültige Bewertung erfolgt anhand des Textes sowie anhand der schriftlich begründeten Einstufungen aus der Erst- und Zweitkorrektur. Da dieses Verfahren besonders aufwändig ist, wird zur Zeit ein ökonomischeres Vorgehen erprobt, bei dem lediglich eine Erst- und eine Zweitkorrektur notwendig sind.

3.3 Probleme bei der Bewertung des Schriftlichen Ausdrucks

Im Gegensatz zum Mündlichen Ausdruck, bei dem anhand von mehreren Items (bis zu 10 Texte) eine Gesamtbewertung vorzunehmen ist, ergeben sich aus dem Format des Schriftlichen Ausdrucks nicht unerhebliche Probleme. Der Schriftliche Ausdruck wird anhand eines einzigen Textes bewertet, denn die Aufgabenstellung verlangt das Abfassen eines zusammenhängenden Textes, der unterschiedliche Schreibhandlungen umfasst. Der zu bewertende Text kann das gesamte von TestDaF erfasste Leistungsspektrum (TDN 3 bis TDN 5) abdecken. Eine möglicherweise schwache Leistung kann daher nicht – wie beim Mündlichen Ausdruck möglich – durch andere Leistungen (andere Texte) kompensiert werden.

Das Testformat ist festgelegt und die geforderten Schreibhandlungen bleiben im Wesentlichen gleich. Auch sind die Aufgaben so konzipiert, dass fachspezifische Kenntnisse bei der Texterstellung nicht erforderlich sind oder Vorteile darstellen. Zudem werden alle Schreibaufgaben ebenso wie alle anderen TestDaF-Aufgaben vor ihrem Einsatz in einer Prüfung erprobt. Dennoch bleiben Phänomene wie kulturelle Geprägtheit von Textsorten bzw. die Übertragung eigenkultureller Textmuster in die schriftliche Produktion in der Fremdsprache und thematische Fairness problematisch. In diesem Zusammenhang ist die Gewährleistung einer gleichbleibenden Schwierigkeit, d.h. hinsichtlich Thema und Aufgabenstellung gleichbleibender Anforderungen, kritisch.⁴

Aus den bisherigen statistischen Analysen der Bewertungsergebnisse wird deutlich, dass die Bewertungsübereinstimmung unbefriedigend ist. Zum Teil liegen Inter-Rater-Übereinstimmungen von deutlich unter 0,5 vor (gewichtetes Kappa von Cohen). Insgesamt verweisen die Daten auf eine mäßige bis schwache Übereinstimmung. Dies stellt zum einen ein Praktikabilitätsproblem dar, denn ca. 50% der schriftlichen Texte müssen eine Drittkorrektur durchlaufen, was u.a. die Ausstellung der Zeugnisse zeitlich verzögern kann. Besonders schwerwiegend ist jedoch das Reliabilitätsproblem, wenn es zu deutlich heterogenen Einschätzungen einer Leistung kommt. Auch eine Intra-Rating-Studie, bei der eine kleine Gruppe von Korrektorinnen in einem gewissen zeitlichen Abstand dieselben Texte erneut bewerteten, zeigte eine unbefriedigende Bewertungsübereinstimmung.

Die Bewertungsarbeit wird u.a. durch Persönlichkeitsfaktoren und typische Effekte beeinflusst. Zu den Persönlichkeitsfaktoren zählen Konzentrationsfähigkeit oder Neigung zu einem eher milden oder eher strengen Urteil. Beispielsweise sollten die Korrektorinnen und Korrektoren bewusst beobachten und individuell entscheiden lernen, in welchem Zeitraum sie wie viele Arbeiten konzentriert und sicher, d.h. anhand gleichbleibender Maßstäbe bewerten können. Besonderes Augenmerk sollte auch auf die Sensibilisierung für die bei der Bewertung auftretenden verzerrenden Effekte gelegt werden. Hierzu zählen Beurteilungsfehler, die durch serielle Effekte, durch Kontaminations- und Halo-Effekte bedingt sind. So beeinflusst unter Umständen eine un-

⁴ Zur Frage der kulturellen Geprägtheit von hochschulrelevanten Textsorten aus Sicht ausländischer Studierender (mexikanische Studierende) siehe z.B. Eßer (1997). Relevant in diesem Zusammenhang ist auch der Vergleich chinesischer und deutscher Aufsatzsorten von Lehker (2001). Zum Problem der inhaltlichen Fairness siehe exemplarisch Hamp-Lyons (1991: 324f.). Einen kurzen Überblick über die Problematik der Bewertung schriftlicher Leistungen geben Alderson/Banerjee (2002: 95-98). Ausführlich wird die Problematik von Weigle (2002) behandelt.

leserliche Handschrift die Bewertung negativ (Charney 1984, zit. in Vaughan 1991: 114). Oder eine nur teilweise schwache Leistung (z.B. die schwache Umsetzung einer Teilaufgabe) beeinflusst die Einschätzung aller übrigen Leistungen negativ. Besondere Bedeutung kommt offensichtlich dem seriellen Effekt (Positionseffekt) zu: Wenn eine Korrektorin bereits eine Reihe von Leistungen als eher schwach eingestuft hat, so tendiert sie bei einer vergleichsweise besseren Leistung unter Umständen dazu, diese als zu hoch einzustufen. Umgekehrt erhält ein relativ schwacher Text in einer Gruppe von sehr guten Leistungen möglicherweise eine zu niedrige Einstufung. Schließlich kann auch die Erwartungshaltung die Bewertung beeinflussen. Liegen Korrekturarbeiten aus einer regionalen Gruppe vor, die gute Leistungen erwarten lässt, so legt der Korrigierende evtl. unbewusst einen eher wohlwollenden oder aber auch einen eher strengen Maßstab an. In Schulungen anhand konkreter Bewertungsarbeiten wird deshalb versucht, das Bewusstsein dafür zu schärfen, dass der TestDaF nicht normorientiert, sondern kriterienorientiert zu bewerten ist.

Die schwachen Bewertungsübereinstimmungen sind vermutlich auch auf die z.T. problematische Gestaltung bzw. Handhabung der Bewertungskriterien (siehe das Bewertungsraster 1 im Anhang) zurückzuführen. Dies zeigen die Analyse von etwa je 100 schriftlich begründeter Bewertungen aus Erst-, Zweit- und Drittkorrektur ebenso wie die eigene Bewertungsarbeit und kritische Rückmeldungen seitens der Korrektorinnen und Korrektoren. Als problematisch angesehen wurde zum einen die sprachliche Gestaltung einiger Deskriptoren sowie ihre logischen Überschneidungen und – damit zusammenhängend – die nicht immer deutliche Trennschärfe zwischen den Deskriptoren, aufgrund derer oftmals eine Entscheidung für die eine oder andere Stufenzuweisung schwer fällt. Ein Grund für diese Schwierigkeit besteht darin, dass die Bewertungsskalen relativ breite Leistungsspektren widerspiegeln. Eine punktgenaue Einordnung ist daher nicht möglich. Bei der Analyse war jedoch auffällig, dass auch die Struktur des Kriterienrasters zu verzerrten Einstufungen führen kann. Dies ist besonders dann der Fall, wenn Leistungen nicht eindeutig einer Leistungsstufe zuzuordnen sind. So wird z.B. der Gesamteindruck anhand der beiden Unterkriterien *Lesefluss* und *Gedankengang* bewertet. Wenn im *Lesefluss* TDN 5 erreicht wird, im *Gedankengang* jedoch TDN 4, dann erfolgt eine Einstufung des Gesamteindrucks auf TDN 5. Diese systematische Aufrundung kann sich bis zur Endbewertung fortsetzen, die anhand der drei übergeordneten Kriterien erfolgt. Somit ergibt sich durch die systematische Aufrundung innerhalb der übergeordneten Kriterien und durch die Errechnung der Endbewertung anhand der drei gleichwertigen Kriterien eine Einstufung, die unter Umständen

der Stufenbeschreibung nicht mehr entspricht. Aus diesen Gründen wurde eine Überarbeitung der Bewertungskriterien des Schriftlichen Ausdrucks notwendig.

3.4 Zur Überarbeitung der Bewertungskriterien des Schriftlichen Ausdrucks

Die Veränderungen betreffen zum einen die sprachliche Gestaltung zum anderen die Struktur des Bewertungsrasters. Im Folgenden soll lediglich auf die Überarbeitung der Struktur näher eingegangen werden (siehe Bewertungsraster 2 im Anhang).

Die Bewertung bzw. Einstufung der gezeigten Leistung erfolgt anhand der drei übergeordneten Kriterien:

- Gesamteindruck,
- Behandlung der Aufgabe,
- Sprachliche Realisierung,

wobei jedoch jedes übergeordnete Kriterium nunmehr aus drei Unterkriterien besteht.

Beim Kriterium *Gesamteindruck* wurde als drittes Unterkriterium der Aspekt *Textstruktur* eingefügt. Der Deskriptor „Text ist klar strukturiert“ (TDN 5) entfällt somit im Kriterium *Behandlung der Aufgabe*. Entsprechend bestehen beide übergeordneten Kriterien aus je drei Unterkriterien. Damit wird der Fokus bei *Behandlung der Aufgabe* auf die inhaltliche Umsetzung der Aufgabe gelegt. Dementsprechend ist die übergeordnete Frage bei *Behandlung der Aufgabe*: „Text wird inhaltlich der Aufgabenstellung gerecht“. Auch beim dritten Kriterium *Sprachliche Realisierung* werden nunmehr drei Unterkriterien angelegt:

- Kohäsion und syntaktische Strukturen,
- Wortschatz,
- Korrektheit.

Nur scheinbar erfährt die *Korrektheit* eine Abwertung und der *Wortschatz* eine Aufwertung. Denn bei den bislang gültigen Kriterien bestand die *Sprachliche Realisierung* aus nur zwei Un-

terkriterien, so dass beispielweise bei einer Bewertung *Sprachliche Mittel* auf der Stufe TDN 4 und *Korrektheit* auf TDN 3 ein Ergebnis von TDN 4 erreicht wird, womit der Aspekt *Korrektheit* gar nicht mehr zum Tragen kommt. Nunmehr besteht eine höhere Wahrscheinlichkeit, dass *Korrektheit* Einfluss nimmt auf die Stufenzuweisung im Kriterium *Sprachliche Realisierung*, denn zumindest zwei der drei Unterkriterien müssen mit TDN 4 bewertet werden, um im Kriterium *Sprachliche Realisierung* TDN 4 zu erreichen.

3.5 Fazit und Forschungsdesiderata

Zusammenfassend kann festgehalten werden, dass die Bewertung des Schriftlichen Ausdrucks in mehrfacher Hinsicht Probleme aufwirft, die untersucht und diskutiert werden müssen:

1. Der Subtest besteht aus lediglich einer Aufgabe mit einem hohen Komplexitätsgrad. Da Bewertungsunterschiede nicht durch Leistungen in anderen Aufgaben kompensiert werden können, führt dies einerseits zu Schwankungen in der Bewertung und andererseits zu geringer Bewertungsübereinstimmung seitens der Korrektorinnen und Korrektoren. Es ist daher zu prüfen, ob das Format zu überarbeiten ist. So könnte über eine größere Zahl von zu produzierenden Texten vermutlich die Reliabilität der Bewertung erhöht werden. Problematisch hierbei wäre jedoch die Augenschein- und Konstruktvalidität, denn im realen Hochschulkontext werden längere zusammenhängende expositorische und argumentative Texte gefordert. Der für TestDaF gewählte Sprachausschnitt (Sprachverwendung im akademischen Kontext) erfordert ein Format, das dieses Konstrukt widerspiegelt (Inhaltsvalidität).
2. Die schwachen Inter-Rater- (und auch Intra-Rater-)Übereinstimmungen verweisen auf typische Faktoren und Effekte, welche die Bewertungsarbeit beeinflussen. Hier sind in Zukunft genauere Analysen anhand von empirischen Daten (z.B. mittels Introspektion bzw. dem Laut-Denken-Verfahren) durchzuführen. Erkenntnisse aus solchen Untersuchungen sollten in die Schulungs- und Kalibrierungsarbeit einfließen. Es gilt, verstärkt Bewusstsein zu schaffen für den Einfluss von Persönlichkeitsfaktoren und sonstigen verzerrenden Effekten bei der Leistungsbewertung.
3. Analysen der Bewertungen und der Bewertungsübereinstimmungen sollen zeigen, ob die überarbeitete Fassung der Bewertungskriterien zu Verbesserungen hinsichtlich Reliabilität, Auswertungsobjektivität und Anwendbarkeit in der Praxis führt.

4. Zu überprüfen bleibt des Weiteren die Validität der Bewertungskriterien (sowohl für den Mündlichen als auch für den Schriftlichen Ausdruck). Hier sollten auch Überlegungen angestellt werden, wie in diesem Kontext am besten methodologisch zu verfahren ist.
5. Mit Hilfe von empirische Langzeitstudien sollte ferner überprüft werden, wie stark die geforderten Leistungen (verbalisiert durch die Deskriptoren) mit den tatsächlichen Anforderungen im Hochschulalltag korrelieren.⁵
6. In diesem Zusammenhang ist auch zu diskutieren, inwiefern biografische Daten der Korrektorinnen und Korrektoren von Bedeutung sind. So ist der berufliche Hintergrund der Korrektorinnen und Korrektoren ausnahmslos der DaF-Bereich. Bei der Auswahl dieses Personenkreises ist gerade die Erfahrung in der Sprachlehre und in der Abnahme von DaF-Prüfungen ein zentrales Kriterium. Einerseits ist hierbei zu vermuten, dass DaF-Lehrkräfte relativ streng syntaktische, morphologische oder orthografische Fehler bewerten. Dieser Aspekt wird dadurch relativiert, dass in den TestDaF-Bewertungsanleitungen die Korrektheit keine zentrale Rolle spielt, sondern die kommunikative Leistung ausschlaggebend ist. Andererseits mögen Aspekte wie „Lesefluss“ und „Gedankengang“ von Personen, die über eine jahrelange Erfahrung mit DaF-Lernenden unterschiedlicher Muttersprache verfügen als eher problemlos eingeschätzt werden, während Personen ohne diese Erfahrungen (z.B. Hochschullehrkräfte aus nicht-philologischen Fachbereichen) möglicherweise eher Rezeptionsschwierigkeiten haben.

4 Kalibrierung der Schwierigkeit der Lese- und Hörverstehensaufgaben anhand von C-Tests als Ankeritems

4.1 Zielsetzung und Vorgehen

Um die Schwierigkeit der Lese- und Hörverstehensaufgaben zu kalibrieren, werden während der Erprobung neuer TestDaF-Aufgaben Ankeritems, d.h. Items mit einem feststehenden Schwierigkeitsgrad eingesetzt und mit Hilfe des Rasch-Modells der probabilistischen Testtheorie skaliert. Bislang wurden zur Verankerung vom *University of Cambridge Local Examinations Syndicate*

⁵ Einen Forschungsüberblick über die für ein Studium an einer englischsprachigen Hochschule geforderten Sprachkenntnisse gibt Johns (1991). Koreik/Mielke (erscheint) haben jüngst in einer empirischen Untersuchung Daten zu den sprachlichen Anforderungen an deutschen Hochschulen erhoben. Es handelt sich um eine Studie, die in erster Linie der Frage nachgeht, ob die sprachliche Vorbereitung durch DSH-Kurs und DSH-Prüfung („Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber“) den tatsächlichen sprachlichen Anforderungen im Verlaufe des Studiums entspricht.

(UCLES) entwickelte Grammatik- und Lexik-Aufgaben im Format kurzer Lückentexte verwendet.

In einer Reihe im Zeitraum Januar 2001 bis April 2002 durchgeführter Untersuchungen wurde überprüft, ob nicht anstelle der bereits mehrfach verwendeten UCLES-Ankeritems C-Tests zur Verankerung neuer TestDaF-Aufgaben eingesetzt werden könnten (vgl. zum Folgenden die ausführliche Darstellung in Arras/Eckes/Grotjahn 2002). Für den Einsatz von C-Tests sprechen u.a. folgende Argumente: (1) Der C-Test eignet sich gut zur Erfassung globaler Sprachkompetenz. (2) Er korreliert in der Regel (relativ) hoch mit dem Lese- und dem Hörverstehen. (3) Er ist sowohl hinsichtlich der Testerstellung als auch hinsichtlich der Testauswertung hoch praktikabel und ökonomisch. (4) Eine computerbasierte Testerstellung, Testdurchführung und Testauswertung ist leicht möglich. Dieses Argument ist auch im Hinblick auf die in Entwicklung befindliche computer-basierte TestDaF-Version zu sehen. (5) Der C-Test ist als Einstufungstest für Deutsch als Fremdsprache an deutschen Hochschulen seit langem etabliert und gilt in diesem Kontext als adäquates Messinstrument.⁶

Bei der Erprobung neuer TestDaF-Aufgaben wurden die von UCLES zur Verfügung gestellten Aufgaben und die eigens entwickelten C-Test-Items zusammen eingesetzt. Den Versuchspersonen war nicht bekannt, dass die UCLES-Aufgaben und der C-Test rein testmethodischen Zielen dienten. Ziel war es, sowohl die neuen Lese- und Hörverstehensaufgaben als auch den C-Test mittels der UCLES-Items zu verankern. Durch die Verankerung des C-Tests sollten die Voraussetzungen geschaffen werden, diesen später selbst als Anker für die Erprobung neuer TestDaF-Aufgaben einsetzen zu können.

4.2 Versuchspersonen

Die Versuchspersonen entsprachen mit Ausnahme einer Voruntersuchung mit Muttersprachlerinnen und Muttersprachlern des Deutschen alle der Zielgruppe von TestDaF: Es handelte sich

⁶ Inwieweit mit Hilfe von C-Tests akademische Sprachfähigkeit im Sinne der *Cognitive-Academic Language Proficiency* von Cummins (1984) erfasst werden kann, hat vor allem Daller (1999) untersucht (vgl. auch Daller/Grotjahn 1999). Weiterführende Hinweise zum C-Test finden sich z.B. in Grotjahn (1995, 2002) oder Grotjahn/Coleman/Raatz (2002) sowie auch unter <http://www.c-test.de>.

um Studierende, die entweder bereits in Deutschland studienvorbereitende Deutschkurse besuchen oder – noch im Heimatland – einen Studienaufenthalt in Deutschland planen.

4.3 Textauswahl

In den verschiedenen Versuchsgruppen wurden unterschiedlich viele C-Test-Texte mit jeweils 20 Lücken eingesetzt. Es wurden dabei Texte aus dem akademischen Kontext ausgewählt, die hinsichtlich Thematik und Textsorte dem TestDaF-Format entsprachen und ein hinreichendes Schwierigkeitsspektrum abdeckten. Die Texte wurden zunächst – der vermuteten Schwierigkeit nach – aufsteigend angeordnet. Die endgültige Auswahl und Anordnung der Texte erfolgte auf der Basis der Ergebnisse der testmethodischen Auswertungen. Zur Bearbeitung standen pro Text fünf Minuten zur Verfügung.

4.4 Testkonstruktion

Die Texte wurden nach dem klassischen Tilgungsprinzip eingerichtet, d.h. beginnend mit dem zweiten Satz wurde die zweite Hälfte jedes zweiten Wortes getilgt. Am Ende des Textes verblieb ein unversehrter (Teil)-Satz. Von der Tilgung ausgenommen blieben Eigennamen und Abkürzungen.

Teils von Beginn an, teils bei der Revision der Items wurde in Ausnahmefällen aus sprachspezifischen Überlegungen abweichend vom Prinzip getilgt. Das betraf vor allem Komposita, deren zweite Komponente nach dem Prinzip $n/2$ bzw. $(n + 1)/2$ entweder ganz oder bis auf einen Buchstaben getilgt worden wäre. Hier war es im Hinblick auf die Lösbarkeit und zur Vermeidung von Varianten notwendig, von der zweiten Komponente genügend Text unversehrt zu lassen.

Das Tilgungsprinzip wurde bei der Revision auch aufgegeben, um die Schwierigkeit von Lücken bzw. Textstellen zu erhöhen, und zwar insbesondere bei Lücken, die bei Deutschlernenden eine sehr hohe Lösungsrate und somit eine geringe Trennschärfe aufwiesen. Weiterhin wurden die Texte auch hinsichtlich der Lexik revidiert, wenn dies zur Vermeidung von akzeptablen Varianten oder zu einer gewünschten Veränderung der Schwierigkeit führte.

4.5 Untersuchungen

In einer ersten Voruntersuchung wurden insgesamt 12 C-Test-Texte in einer Gruppe von deutschen Muttersprachlern eingesetzt. Nach weiteren Analysen und Erprobungen mit Deutschlernern wurde schließlich eine Auswahl von vier Texten vorgenommen. Diese Texte wurden dann im Rahmen der Erprobung zweier neuer TestDaF-Prüfungssätze zusammen mit den von UCLES entwickelten Ankeritems überprüft: Zunächst jeweils im Rahmen von Vorerprobungen mit 20 bzw. 12 Personen mit Deutsch als Muttersprache, anschließend weltweit mit 187 bzw. 144 Deutschlernenden. Im Frühjahr 2002 wurden die C-Tests, wiederum zusammen mit den UCLES-Ankeritems, erneut im Rahmen dreier Erprobungssätze eingesetzt.

4.6 Auswertungskategorien

Da nicht ausgeschlossen werden konnte, dass bei der TestDaF-Zielgruppe unterschiedliche Modalitäten bei der C-Test-Auswertung zu Unterschieden in der Rangordnung der Probanden führen, erschien es notwendig, bei der Entwicklung der Ankeritems zunächst möglichst detaillierte Auswertungskategorien zu verwenden. Zudem erlaubt eine Auswertung anhand detaillierter Kategorien ein besseres Verständnis der von den Probanden jeweils erbrachten Leistung. Allerdings wird dieser Vorteil mit einer gewissen Unschärfe bei der Zuordnung zu den Kategorien und zugleich mit einem erheblichen Verlust an Auswertungsökonomie erkaufte (das Problem der Auswertungskategorien wird ausführlich in Arras/Eckes/Grotjahn 2002 diskutiert).

Im Rahmen der Vorerprobung an Muttersprachlerinnen und Muttersprachlern und im Rahmen der Erprobung an DaF-Gruppen in Bochum und Münster sowie bei der Erprobung neuer TestDaF-Aufgaben (Stichprobe E004) wurde zwischen folgenden sechs Auswertungskategorien unterschieden:⁷

Kategorie 1: unausgefüllt

Kategorie 2: orthografisch richtiges Original

Kategorie 3: grammatisch und/oder inhaltlich nicht akzeptabel

⁷ Andere, ökonomischere Kodierungen sehen lediglich die Kategorien „nicht ausgefüllt“, „falsch ausgefüllt“, „richtig ausgefüllt“ vor, wobei zur Kategorie „falsch ausgefüllt“ sowohl grammatische, lexikalische als auch orthografische Fehler gezählt werden.

Kategorie 4: orthografisch richtige Variante

Kategorie 5: orthografisch falsches Original

Kategorie 6: orthografisch falsche Variante

Die Kategorie „grammatisch und/oder inhaltlich nicht akzeptabel“ erweist sich zwar als ökonomisch bei der raschen Testauswertung, doch verhindert sie eine genauere Fehlerzuordnung und damit auch eine genauere Fehleranalyse sowie eine exaktere Einschätzung der Fähigkeit, da sie zwei unterschiedliche Bereiche umfasst: Inhalt und Grammatik, wobei Grammatik selbst unterschiedliche Teilaspekte beinhaltet (Syntax, Morphologie).

Um Fehler präziser zuordnen zu können, wurde daher die Auswertungskategorie „grammatisch und/oder inhaltlich nicht akzeptabel“ folgendermaßen aufgelöst:

Kategorie 3 (neu): inhaltlich nicht akzeptabel

Kategorie 7: inhaltlich akzeptabel, aber grammatisch falsch

Auf diese Weise war es möglich, zwischen Versuchspersonen zu unterscheiden, die auf der Makroebene vermutlich über die erforderliche Textverstehenskompetenz verfügen und deshalb lexikalisch richtig, jedoch z.B. morphologisch falsch ergänzen, und solchen Versuchspersonen, die lexikalisch kontextuell unangemessen ergänzen. Ab der Erprobung E005 wurden bei der Auswertung der C-Tests daher sieben Kategorien angelegt. Diese Daten wurden allerdings bisher noch nicht im Detail ausgewertet. Die folgenden Aussagen beziehen sich deshalb stets auf das ursprüngliche sechsstufige Kategoriensystem.

4.7 Statistische Analysen

Zunächst wurde überprüft, inwieweit das Auswertungsverfahren – und zwar speziell die unterschiedliche Kategorisierung von orthografischen Varianten⁸ – einen Einfluss auf die Punktwerte und die Rangfolge der Probanden sowie auf die Reliabilität und Validität des C-Tests hat und ob letztendlich nicht eine ökonomischere Auswertung vertretbar ist (anhand von zwei Gruppen von Lernern: $N = 93$ und $N = 187$).

⁸ Beim praktischen Einsatz von C-Tests stellt sich immer wieder die Frage, wie insbesondere Orthografiefehler bei der Auswertung zu behandeln sind.

In weiteren Analysen wurde überprüft, inwieweit die vier C-Test-Texte auf der Basis des Rasch-Modells – und zwar des kontinuierlichen Ratingskalen-Modells von Müller (1999) – skalierbar sind und damit der vorliegende C-Test als eindimensionales Messinstrument angesehen werden kann.

Schließlich wurden die Reliabilität des C-Tests und der TestDaF-Subtests sowie die Interkorrelationen zwischen C-Test und den vier TestDaF-Subtests ermittelt. Weiterhin wurde überprüft, ob sich der C-Test möglicherweise für ein Screening schwächerer Kandidaten eignet.

Die durchgeführten Untersuchungen haben folgende Hauptergebnisse im Hinblick auf die Entwicklung von deutschen C-Tests und den Einsatz von C-Tests im Rahmen von TestDaF erbracht:

1. Es sollten möglichst sowohl Originallösungen als auch akzeptable Varianten als korrekt gewertet werden. Bei nicht sehr weit fortgeschrittenen Lernern sind akzeptable Varianten allerdings relativ selten und Bewertungsunterschiede haben deshalb nur einen geringen Effekt.
2. Gehört die orthografische Kompetenz mit zum messenden Konstrukt, sind orthografisch falsche Originale und Varianten als falsche Lösung zu werten.
3. Orthografiefehler kommen erwartungsgemäß vor allem bei weniger fortgeschrittenen Lernern vor. Hier kann es je nach Wertung von Orthografiefehlern zu deutlich unterschiedlichen Punktwerten und Rangfolgen der Probanden kommen.
4. Durch gründliche Vorerprobung und Revision lassen sich C-Tests entwickeln, in denen nur wenige Orthografiefehler vorkommen und die Wertung von Orthografiefehlern als korrekt kaum Auswirkungen auf die Gesamtpunktwerte hat. In einem solchen Fall kann es aus Ökonomiegründen gerechtfertigt sein, orthografisch falsche Lösungen als „nicht korrekt“ zu werten, auch wenn die orthografische Kompetenz nicht zum messenden Konstrukt gehört. Allerdings kann auch bei einer nur relativ kleinen Zahl von orthografisch falschen Lösungen die Wertung von Orthografiefehlern als „nicht korrekt“ zu beträchtlichen Rangplatzdifferenzen führen.
5. Der im Kontext von TestDaF entwickelte, aus vier Texten mit jeweils 20 Lücken bestehende C-Test ist hoch reliabel (Cronbach's Alpha = 0,84; $N = 187$). Die Analysen auf der Basis des kontinuierlichen Ratingskalen-Modells von Müller (1999) belegen zudem die gelungene Rasch-Skalierung der vier Texte ($N = 187$).
6. Der C-Test korreliert relativ hoch mit allen vier Subtests des TestDaF (Spearman's r zwischen 0,64 und 0,68; N zwischen 113 und 187). Bemerkenswert ist insbesondere die hohe Korrelation von 0,64 mit dem Mündlichen Ausdruck.

7. Der C-Test scheint sich sowohl für die Kalibrierung der Lese- und Hörverstehensitems als auch für ein ökonomisches Screening schwächerer Probanden zu eignen. Ein solches Screening bietet sich vor allem bei den hinsichtlich der Auswertung aufwändigen Prüfungsteilen Schriftlicher und Mündlicher Ausdruck, an – z.B. auf der Basis einer freiwilligen Selbstevaluation mit Hilfe eines web-basierten, weltweit einsetzbaren und automatisch auswertbaren C-Tests.

5 Weitere Entwicklungen

Das TestDaF-Institut entwickelt derzeit in Zusammenarbeit mit dem Fachbereich Informatik an der FernUniversität Hagen eine computer-basierte Version des TestDaF (CBT-Version).⁹ Ziel ist, die Prüfung in Zukunft häufiger und ggf. zu individualisierten Testterminen anbieten zu können. Die CBT-Vorversion orientiert sich stark am Paper-and-Pencil-Test (siehe Modellsatz, der über die Homepage <http://www.testdaf.de> zugänglich ist). Zum einen sind Fragen wie z.B. technische Ausstattung der Testzentren oder das Problem länderspezifischer Tastaturbelegungen bislang nicht hinreichend geklärt. Inwiefern eine inhaltliche Veränderung des Formats (Aufgabenstellungen, Item-Anzahl, Textlängen, Präsentation u.Ä.) vorgenommen werden muss, wird eine erste Erprobungsphase zeigen. Eine Vorversion des CBT wird voraussichtlich noch im Jahre 2002 erprobt und soll Aufschluss hinsichtlich der aufgeworfenen Fragen bringen.

⁹ Erste Überlegungen hierzu wurden bereits in der Entwicklungsphase des TestDaF angestellt (vgl. Gutzat/Pauen/Voss 2000; Six 2000).

6 Literaturverzeichnis

- Alderson, J. Charles/Banerjee, Jayanti (2002):** "Language testing and assessment (Part 2)". *Language Teaching* 35(1), 79-113.
- Arras, Ulrike/Eckes, Thomas/Grotjahn, Rüdiger (2002):** „C-Tests im Rahmen des ‚Test Deutsch als Fremdsprache‘ (TestDaF): Erste Forschungsergebnisse“. In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen: Bd. 4*. Bochum: AKS-Verlag, 175-209.
- Arras, Ulrike/Grotjahn, Rüdiger (erscheint):** „TestDaF: Einige aktuelle Entwicklungen“. In: Dokumentation der 22. Arbeitsstagung des AKS in Chemnitz, Februar 2002.
- Bolton, Sibylle (Hrsg.) (2000):** *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. Köln: Gilde.
- Charney, D. (1984):** The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English* 18, 65-81.
- Coleman, James A./Grotjahn, Rüdiger/Raatz, Ulrich (eds.) (2002):** *University language testing and the C-Test*. Bochum: AKS-Verlag.
- Cummins, Jim. (1984):** "Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students". In: Rivera, Charlene (ed.), *Language proficiency and academic achievement*. Clevedon, England: Multilingual Matters, 2-19.
- Daller, Helmut (1999):** *Migration und Mehrsprachigkeit: Der Sprachstand türkischer Rückkehrer aus Deutschland*. Frankfurt am Main: Lang.
- Daller, Helmut/Grotjahn, Rüdiger (1999):** "The language proficiency of Turkish returnees from Germany: An empirical investigation of academic and everyday language proficiency". *Language, Culture and Curriculum* 12(2), 156-172.
- Eßer, Ruth (1997):** 'Etwas ist mir geheim geblieben am deutschen Referat'. *Kulturelle Geprägtheit wissenschaftlicher Textproduktion und ihre Konsequenzen für den universitären Unterricht von Deutsch als Fremdsprache*. München: Iudicium.
- Europarat (2001):** *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin: Langenscheidt.
- Grotjahn, Rüdiger (1995):** „Der C-Test: State of the Art“. *Zeitschrift für Fremdsprachenforschung* 6(2), 37-60.
- Grotjahn, Rüdiger (erscheint):** „TestDaF: Theoretical Basis and Empirical Research“. Paper presented at the ALTE European Year of Languages Conference, 5 to 7 July 2001, Barcelona. [erscheint 2002 in den Tagungsberichten]
- Grotjahn, Rüdiger (Hrsg.) (2002):** *Der C-Test. Theoretische Grundlagen und praktische Anwendungen: Bd. 4*. Bochum: AKS-Verlag.

- Grotjahn, Rüdiger/Kleppin, Karin (2001):** „TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis“. In: Aguado, Karin/Riemer, Claudia (Hrsg.), *Wege und Ziele. Zur Theorie, Empirie und Praxis des Deutschen als Fremdsprache (und anderer Fremdsprachen)*. Festschrift für Gert Henrici zum 60.Geburtstag. Baltmannsweiler: Schneider Verlag Hohengehren, 419-433.
- Gutzat, Bärbel/Pauen, Peter/Voss, Josef (erscheint):** „Computer- und Interneteinsatz bei TestDaF“. In: Dokumentation der 21. Arbeitstagung des AKS am 10.3.2000 in Saarbrücken [abrufbar unter: <http://www.testdaf.de>].
- Hamp-Lyons, Liz (1991):** “Issues and Directions in Assessing Second Language Writing in Academic Contexts”. In: Hamp-Lyons, Liz (ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood: Ablex, 323-329.
- Johns, Ann (1991):** “Faculty Assessment of ESL Student Literacy Skills: Implications for Writing Assessment“. In: Hamp-Lyons, Liz (ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood: Ablex, 167-179.
- Kniffka, Gabriele/Üstünsöz-Beurer, Dörthe (2001):** „TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassettengesteuerten Testformats“. *Fremdsprachen Lehren und Lernen* 30, 127-149.
- Koreik, Uwe/Mielke, Frank (erscheint):** „DSH-Kurse auf dem Prüfstand“. In: Voss, Bernd (Hrsg.), *Fremdsprachenunterricht auf dem Prüfstand, Innovation – Qualität – Evaluation. Beiträge zum 19. Kongress für Fremdsprachendidaktik der Deutschen Gesellschaft für Fremdsprachenforschung in Dresden*. Berlin: Pädagogischer Zeitschriftenverlag.
- Lehker, Marianne (2001):** „Chinesische und deutsche Aufsatzsorten im Vergleich“. In: Fix, Ulla/Habscheid, Stephan/Klein, Josef (Hrsg.), *Zur Kulturspezifik von Textsorten*. Tübingen: Stauffenburg, 131-146.
- Müller, Hans (1999):** *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen: Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items*. Bern: Huber.
- Projektgruppe TestDaF (2000):** „TestDaF: Konzeption, Stand der Entwicklung, Perspektiven“. *Zeitschrift für Fremdsprachenforschung* 11(1), 63-82.
- Six, Hans-Werner (2000):** „Informatikunterstützung für TestDaF“. In: Bolton, Sibylle (Hrsg.), *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. Köln: Gilde, 131-140.
- Vaughan, Caroline (1991):** “Holistic Assessment: What Goes On in the Rater’s Mind?”. In: Hamp-Lyons, Liz (ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood: Ablex, 111-125.
- Weigle, Sarah C. (2002):** *Assessing Writing*. Cambridge: Cambridge University Press.

7 Ausgewählte Informationsmaterialien zu TestDaF
(z.T. in englischer, französischer, spanischer, russischer und chinesischer Sprache)

TestDaF im Überblick, TestDaF-Institut Hagen 2001

Faltblatt Informationen zu TestDaF, TestDaF-Institut Hagen 2001

Kurzbeschreibung der Subtests, TestDaF-Institut Hagen 2001

Modellsatz 01, TestDaF-Institut Hagen 2001

zugänglich über die Website <http://www.testdaf.de>

oder erhältlich über die Kontaktadresse info@testdaf.de
TestDaF-Institut
Elberfelder Str. 103
58084 Hagen

Anhang

Schriftlicher Ausdruck Modellsatz 01

Bewertungskriterien zum Schriftlichen Ausdruck

Bewertungskriterien zum Schriftlichen Ausdruck, überarbeitete Fassung

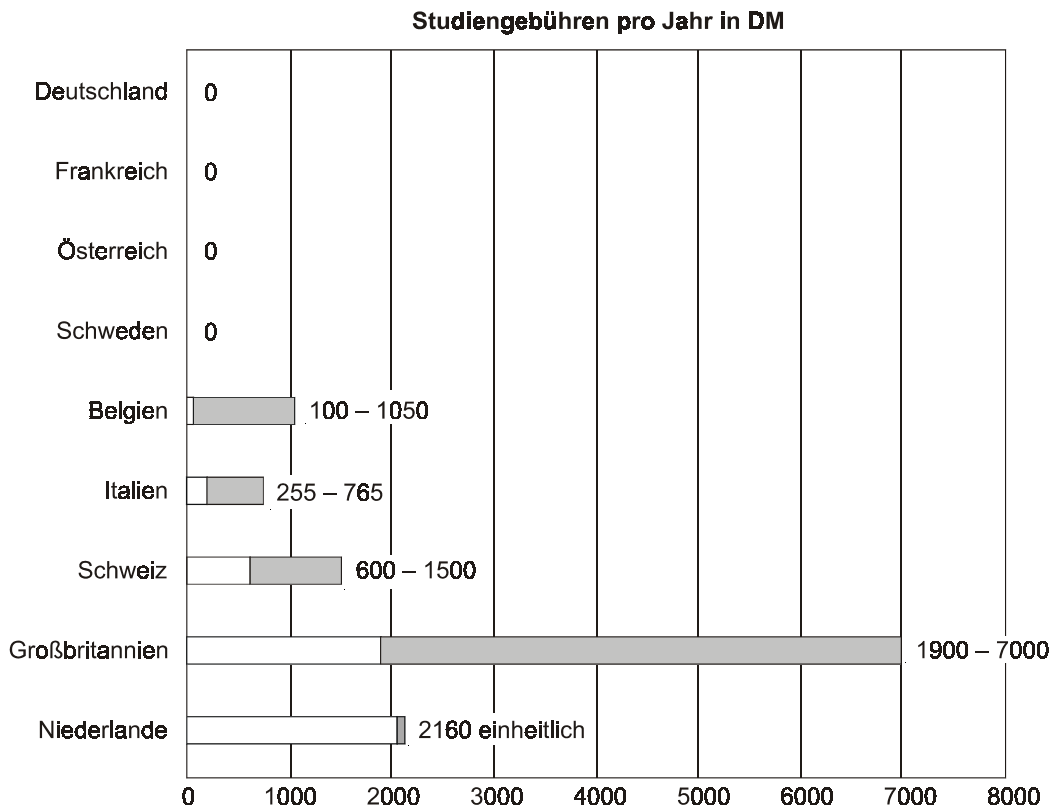
Thema

Studiengebühren

Bis in die 60er Jahre mussten die Studenten in Deutschland Studiengebühren zahlen. Dann wurden die Studiengebühren im Zusammenhang mit Universitätsreformen abgeschafft, denn alle sollten gleiche Bildungschancen haben. Jetzt wird in Deutschland diskutiert, ob Studiengebühren wieder eingeführt werden sollen.



Schreiben Sie einen Text zum Thema „Studiengebühren“



Quelle: Forschung und Lehre 8/1999, S. 398

Beschreiben Sie, wie die Frage von Studiengebühren in verschiedenen europäischen Ländern geregelt ist: Vergleichen Sie

- die Höhe der Studiengebühren.
- die Bandbreite der Studiengebühren, d.h. die Unterschiede innerhalb einzelner Länder.

In der Diskussion um Studiengebühren werden zwei Meinungen vertreten:

- **Wenn die Hochschulen Studiengebühren verlangen, dann können sie das Studienangebot verbessern und die Studenten sind motivierter, gute Leistungen zu erbringen.**
- **Bildung ist eine Aufgabe des Staates, deshalb darf es keine Studiengebühr geben.**
 - Geben Sie die beiden Aussagen mit eigenen Worten wieder.
 - Nehmen Sie zu beiden Aussagen Stellung und begründen Sie Ihre Stellungnahme.
 - Gehen Sie auch auf die Situation in Ihrem Heimatland ein.

Schriftlicher Ausdruck Bewertungskriterien

	TDN 5	TDN 4	TDN 3	unter TDN 3
Gesamteindruck	Der Text liest sich durchgängig flüssig.	An einzelnen Stellen gerät der Lesefluss ins Stocken.	An manchen Stellen ist wiederholtes Lesen erforderlich.	Der Text liest sich insgesamt nicht flüssig.
	Der Gedankengang kann problemlos nachvollzogen werden.	Der Gedankengang kann nachvollzogen werden, wenn auch vereinzelt die Rezeption verzögert wird.	Der Gedankengang kann von einem kooperativen Leser nachvollzogen werden.	Der Gedankengang kann nur mühsam oder bruchstückhaft nachvollzogen werden.
Behandlung der Aufgabe	<i>Der Text wird der Aufgabenstellung gerecht:</i>	<i>Der Text wird der Aufgabenstellung weitgehend gerecht:</i>	<i>Der Text wird der Aufgabenstellung noch gerecht:</i>	<i>Der Text wird der Aufgabenstellung nicht gerecht:</i>
	Der Text ist klar strukturiert.	Der Text ist insgesamt klar strukturiert.	Der Text weist vereinzelt gedankliche Sprünge auf.	Der Text ist nicht klar strukturiert.
	Alle in der Aufgabenstellung genannten Punkte werden in ausreichendem Umfang behandelt.	Alle in der Aufgabenstellung genannten Punkte werden behandelt, manche jedoch zu knapp.	Fast alle in der Aufgabenstellung genannten Punkte werden behandelt.	Nur einige in der Aufgabenstellung genannte Punkte werden behandelt.
	Die Informationen der Grafik werden zusammengefasst; sie werden klar und folgerichtig dargestellt.	Die wesentlichen Informationen der Grafik werden klar und folgerichtig wiedergegeben.	Die Informationen der Grafik werden aufzählend wiedergegeben.	Die Beschreibung der Grafik ist ohne Kenntnis der Vorgabe nicht verständlich.
	Die Argumente werden sachlich und ausführlich genug begründet und ggfs. mit Beispielen belegt.	Die Argumente werden sachlich begründet, manchmal nur knapp, und ggfs. mit Beispielen belegt.	Der eigene Standpunkt wird vorgebracht und durch persönliche Wertungen verstärkt.	Der eigene Standpunkt wird nicht oder nur in Ansätzen verdeutlicht.

<p style="text-align: center;">Sprachliche Realisierung</p>	<p>Die Wahl der sprachlichen Mittel ist der Aufgabenstellung angemessen:</p> <ul style="list-style-type: none"> • breites Spektrum an kohäsionsstiftenden Mitteln • breites Spektrum an syntaktischen Strukturen • weitgehend differenzierter und präziser Wortschatz <p>Der Text enthält vereinzelt morphosyntaktische, lexikalische und orthografische Fehler.</p>	<p>Die Wahl der sprachlichen Mittel ist der Aufgabenstellung insgesamt angemessen:</p> <ul style="list-style-type: none"> • begrenztes Spektrum an kohäsionsstiftenden Mitteln • syntaktische Strukturen variieren • breiter Wortschatz <p>Der Text enthält gelegentlich (nicht-systematische) morphosyntaktische, lexikalische und orthografische Fehler.</p>	<p>Die Wahl der sprachlichen Mittel ist der Aufgabenstellung nicht immer angemessen:</p> <ul style="list-style-type: none"> • einfache Verknüpfungselemente • einige Variationen bei den syntaktischen Strukturen • ausreichender Wortschatz <p>Der Text enthält morphosyntaktische, lexikalische und orthografische Fehler, die das Verstehen manchmal beeinträchtigen.</p>	<p>Die Wahl der sprachlichen Mittel ist der Aufgabenstellung nicht angemessen:</p> <ul style="list-style-type: none"> • kaum Verknüpfungselemente • nur wenige Variationen bei den syntaktischen Strukturen • eingeschränkter Wortschatz <p>Der Text enthält morphosyntaktische, lexikalische und orthografische Fehler, die das Verstehen deutlich erschweren.</p>
--	---	---	---	--

Schriftlicher Ausdruck Bewertungskriterien – überarbeitete Fassung

	TDN 5	TDN 4	TDN 3	unter TDN 3
Gesamtein- druck	<p>1. Der Text liest sich durchgängig flüssig.</p> <p>2. Der Gedankengang kann problemlos nachvollzogen werden.</p> <p>3. Der Text ist klar strukturiert.</p>	<p>1. An einzelnen Stellen gerät der Lesefluss ins Stocken.</p> <p>2. Der Gedankengang kann nachvollzogen werden, wenn auch vereinzelt die Rezeption verzögert wird.</p> <p>3. Der Text ist insgesamt noch strukturiert.</p>	<p>1. An manchen Stellen ist wiederholtes Lesen erforderlich.</p> <p>2. Der Gedankengang kann von einem kooperativen Leser nachvollzogen werden.</p> <p>3. Der Text weist Brüche auf.</p>	<p>1. Der Text liest sich insgesamt nicht flüssig.</p> <p>2. Der Gedankengang kann nur mühsam oder bruchstückhaft nachvollzogen werden.</p> <p>3. Der Text ist nicht klar strukturiert.</p>
Behandlung der Aufgabe	<p><i>Der Text wird der Aufgabenstellung inhaltlich gerecht:</i></p> <p>1. Alle in der Aufgabenstellung genannten Punkte werden in ausreichendem Umfang behandelt.</p> <p>2. Die Informationen der Grafik(en) werden zusammengefasst; sie werden klar und folgerichtig dargestellt.</p> <p>3. Im argumentativen Teil wird sachlich und ausführlich genug begründet und ggf. werden Beispiele als Belege angeführt.</p>	<p><i>Der Text wird der Aufgabenstellung inhaltlich weitgehend gerecht:</i></p> <p>1. Alle in der Aufgabenstellung genannten Punkte werden behandelt, manche jedoch zu knapp.</p> <p>2. Die Informationen der Grafik(en) werden klar und folgerichtig wiedergegeben.</p> <p>3. Im argumentativen Teil wird sachlich begründet, z.T. nur knapp, und ggf. werden Beispiele als Belege angeführt.</p>	<p><i>Der Text wird der Aufgabenstellung inhaltlich noch gerecht:</i></p> <p>1. Fast alle in der Aufgabenstellung genannten Punkte werden behandelt.</p> <p>2. Die Informationen der Grafik(en) werden überwiegend aufzählend wiedergegeben.</p> <p>3. Im argumentativen Teil werden Standpunkte/ Überlegungen deutlich und ggf. durch persönliche Wertungen verstärkt.</p>	<p><i>Der Text wird der Aufgabenstellung inhaltlich nicht gerecht:</i></p> <p>1. Nur einige in der Aufgabenstellung genannte Punkte werden behandelt.</p> <p>2. Die Beschreibung der Grafik(en) ist nicht verständlich.</p> <p>3. Im argumentativen Teil werden Standpunkte/ Überlegungen nicht oder nur in Ansätzen verdeutlicht.</p>

<p>Sprachliche Realisierung</p>	<p><i>Die sprachliche Realisierung ist der Aufgabenstellung angemessen:</i></p> <ol style="list-style-type: none"> 1. Der Text hat <ul style="list-style-type: none"> - ein breites Spektrum an kohäsionsstiftenden Mitteln - ein breites Spektrum an syntaktischen Strukturen. 2. Der Wortschatz ist weitgehend differenziert und präzise. 3. Der Text enthält vereinzelt morphosyntaktische, lexikalische und orthografische Fehler. 	<p><i>Die sprachliche Realisierung ist der Aufgabenstellung weitgehend angemessen:</i></p> <ol style="list-style-type: none"> 1. Der Text hat <ul style="list-style-type: none"> - ein begrenztes Spektrum an kohäsionsstiftenden Mitteln - ein begrenztes Spektrum an syntaktischen Strukturen. 2. Der Wortschatz ist breit, teilweise jedoch nicht präzise. 3. Der Text enthält gelegentlich (nicht-systematische) morphosyntaktische, lexikalische und orthografische Fehler, die das Verstehen jedoch nicht beeinträchtigen. 	<p><i>Die sprachliche Realisierung ist der Aufgabenstellung nicht immer angemessen:</i></p> <ol style="list-style-type: none"> 1. Der Text hat <ul style="list-style-type: none"> - einfache Verknüpfungselemente - einige Variationen bei den syntaktischen Strukturen. 2. Der Wortschatz ist ausreichend. 3. Der Text enthält morphosyntaktische, lexikalische und orthografische Fehler, die das Verstehen beeinträchtigen. 	<p><i>Die sprachliche Realisierung ist der Aufgabenstellung nicht angemessen:</i></p> <ol style="list-style-type: none"> 1. Der Text hat <ul style="list-style-type: none"> - kaum Verknüpfungselemente - nur wenige Variationen bei den syntaktischen Strukturen. 2. Der Wortschatz ist eingeschränkt. 3. Der Text enthält morphosyntaktische, lexikalische und orthografische Fehler, die das Verstehen deutlich erschweren.
--	--	--	--	--