Die Frankfurter Aufnahmeprüfung

Heinrich Stalb

1. Geschichte

In den Anfangsjahren des Studienkollegs Frankfurt, d. h. in den sechziger Jahren, waren der Unterricht und insbesondere die Prüfungen fast ausschließlich an der muttersprachlichen Didaktik orientiert. Wie beim damaligen Übergang von der Grundschule zum Gymnasium bestand daher zum Beispiel die Aufnahmeprüfung am Studienkolleg aus einem Diktat und einem kleinen Aufsatz. Das Fach Deutsch wurde nicht, jedenfalls nicht prinzipiell, als Deutsch als Fremdsprache verstanden. Folglich blieben Untersuchungen zu Tests und Prüfungen im Fremdsprachenunterricht völlig unbeachtet, z. B. die grundlegenden Arbeiten von Robert Lado oder Rebecca M. Valette, die im Original immerhin seit 1961 bzw. 1967 vorlagen.

Erst Anfang der siebziger Jahre wurden erste fremdsprachendidaktische Prüfungsformen übernommen. Man ließ jetzt Lückentexte zur Erfassung von Grammatik und Wortschatz ausfüllen, Fragen zu gehörten Texten beantworten und Bildbeschreibungen zur Überprüfung der Schreibfertigkeit anfertigen. Dabei war der Versuch, auch das Hörverstehen als eigene Fertigkeit zu erfassen, selbst im Fremdsprachenunterricht der deutschen Schulen damals noch Neuland.

Die neue Aufnahmeprüfung war also zweifellos ein ganz erheblicher Fortschritt. Dennoch zeigten sich bald drei gravierende Probleme. Weil für jede Prüfung neue Texte und Aufgaben gesucht werden mußten, schwankte das Anspruchsniveau erheblich. Zweitens: Den Bildbeschreibungen lagen die bekannten Vater-und-Sohn-Geschichten zugrunde (E.O. Plauen 1964). Die Erwartung war, daß ihr Humor den Prüfungsstreß reduzieren, vor allem aber, daß die Darstellung einer Bild-Geschichte die üblichen Beschreibungsversatzstücke (oben, links, ist, befindet sich, sieht man etc.) zumindest ansatzweise durch freiere Schreibäußerungen ergänzen würde. Nicht bedacht wurde, daß auch das Bildverstehen ganz entscheidend von fremdkulturellen Vorerfahrungen bestimmt wird. So zeigte sich, daß die Bildinformationen manchmal gar nicht als Darstellung einer Geschichte erkannt oder für die Korrektoren in so fremdartiger Weise interpretiert wurden, daß die Bewertung erheblich beeinträchtigt war. Drittens schließlich wurde als Mangel empfunden, daß die Lesefertigkeit beim Aufnahmeprüfungsverfahren völlig unberücksichtigt blieb.

Es gab also Gründe genug, parallel und alternativ zum eigenen Prüfungsmodus fremde Konzepte zu erproben. Das geschah einige Male, und zwar durch den Rückgriff auf das Zertifikat Deutsch als Fremdsprache. Die Tests zur Erlangung des Zertifikats wurden entwickelt vom Deutschen Volkshochschul-Verband, dem Goethe-Institut und der Universität Freiburg (*Das Zertifikat* 1985: 4). Sie überprüften alle sprachlichen Fertigkeiten, also auch das Leseverständnis, erfaßten

die schriftliche Ausdrucksfähigkeit ohne Bildstimuli und stammten aus einer Item-Bank, die durch Inhalte und Testformate ein standardisiertes Anspruchsniveau sicherte. Obwohl das Zertifikat dem fremdsprachendidaktischen Wissensstand der damaligen Zeit voll entsprach und bis zu seiner Revision 1998/99 weltweit mit Erfolg eingesetzt wurde, konnte es sich am Studienkolleg, nicht nur in Frankfurt, nicht durchsetzen. Dafür gab es im wesentlichen zwei Gründe. Die Zertifikats-Prüfungen waren erstens zu leicht. Die unterste Bestehensgrenze betrug 60% des maximal erreichbaren Punktwertes. Nach den Erfahrungen in Frankfurt hätten wir mindestens 75% ansetzen müssen. Das hätte die mißliche Folge gehabt, den Bewerbern mit 60 bis 74% der Gesamtpunktzahl vermitteln zu müssen, daß sie zwar die Prüfung bestanden haben, aber trotzdem am Kolleg nicht aufgenommen werden können. Zweitens war das Zertifikat zu einseitig an der Sprachkompetenz orientiert, die in alltagssprachlichen Situationen erforderlich ist. Die Zertifikatslernziele waren, wenn auch in neueren Fassungen etwas weniger betont, auf die sprachlichen Bedürfnisse von Touristen ausgerichtet (*Das Zertifikat* 1985: 12), nicht jedoch auf Studienbewerber, die sich auch im Fachunterricht (Biologie, Chemie, Soziologie etc.) bewähren sollten. Die Studienkollegs mußten also ihre eigene Aufnahmeprüfung noch finden.

2. Gegenwart

2.1 Beschreibung der Prüfung

Ende der siebziger, Anfang der achtziger Jahre nahm die Frankfurter Aufnahmeprüfung schrittweise ihre heutige Form an. Sie wurde seitdem durch neue Texte vervollständigt, in einigen Textinhalten aktualisiert, in vielen Items mit dem Blick auf Schwierigkeit, Trennschärfe, Anforderungen der kommunikativen Kompetenz revidiert und im Umfang so ergänzt, daß inzwischen acht Prüfungssätze zur Verfügung stehen. Jeder Prüfungssatz besteht aus drei Teilen: einem Teil zur Überprüfung von Grammatik und Wortschatz, einem zur Überprüfung des Hörverständnisses und einem zur Überprüfung der Schreibfertigkeit.

Grammatik und Wortschatz werden in Form von Lückenaufgaben erfaßt. Dabei werden maximal 50 Punkte vergeben. Das Hörverständnis wird durch Ja-/Nein-Aufgaben kontrolliert. Die hier erreichbare Punktzahl unterscheidet sich in den einzelnen Prüfungssätzen; sie liegt etwa bei 35 Punkten. Beim Schreibtest soll ein von einem Prüfer vorgelesener und dabei gleichzeitig von den Prüflingen mitgelesener Text wiedergegeben werden. Bei der Bewertung werden Punkte für die Richtigkeit und Vollständigkeit des wiedergegebenen Inhalts und Punkte für die sprachliche Korrektheit vergeben. Die sprachliche Korrektheit wird doppelt gewichtet. Die maximal erreichbare Punktzahl in diesem Prüfungsteil liegt bei 60 Punkten. Die Bestehensgrenze wurde grundsätzlich auf 55 Prozent der maximal erreichbaren Punktzahl festgelegt. Je nach Bewerberzahl können von den Interessenten für G-Kurse (= Geisteswissenschaften) gelegentlich 60 Punkte verlangt werden. Bei Bewerbern für andere Kurse (M = Medizin, W = Wirtschaft, N = Naturwissenschaften/Technik) müssen wir die Ansprüche öfter auf 50 Prozent senken. Weitere Einzelheiten zur Aufnahmeprüfung sind der folgenden Übersicht zu entnehmen.

2.2 Grundlagen

Ein wichtiger Ausgangspunkt für die Frankfurter Prüfung waren Lados Test- und Sprachvorstellungen, nach denen Elemente (Komponenten) wie Aussprache, Morphologie und Syntax, Wort-

Die Frankfurter Aufnahmeprüfung im Überblick

Prüfungsteil	Aufgabe (Verfahren)	Punkte	Zeit
Strukturtest: Wortschatz / Grammatik	In 4 (von 8) Prüfungssätzen: 50 Einzelsätze (teilweise Minidialoge) mit Lücken; in den 4 anderen Prüfungssätzen: 25 Einzelsätze (teilweise Minidialoge) und ein fortlaufender Text mit 25 Lücken	50	35 Minuten Arbeitszeit
Hörtest	Zu einem zweimal vorgelesenen Text rund 35 Aussagen mit drei Antwortmöglichkeiten: ja, nein, weiß nicht; mit Ausnahme von 2 Prüfungssätzen auch Erfassung von Globalverständnis / tieferem Verständnis	rund 35	20 Minuten Arbeitszeit
Pause			30 Minuten
Schreibtest	Wiedergabe eines zweimal vorgelesenen und von den Prüflingen mitgelesenen Texts; bewertet nach Inhalt (1/3) und Sprache (2/3); Wiedergabehilfe: Wörterliste	60	55 Minuten Arbeitszeit

schatz, kulturelle Inhalte, die in Fertigkeiten wie Sprechen, Hören, Lesen, Schreiben integriert sind, Aufschluß über den Grad der Sprachbeherrschung geben (Lado 1971: 40 f.). Lados Vorstellungen sind im Laufe der Jahre von unterschiedlichen Positionen her kritisiert worden. Morrow z. B. wirft Lado eine atomistische Testkonzeption vor, die davon ausgehe, daß das Beherrschen der Elemente einer Sprache dasselbe wie das Beherrschen der Sprache sei.

»Knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language.« (Morrow 1979: 145)

Die Kenntnis der Elemente einer Sprache wird hier stark abgewertet. Dabei wird aber etwas sehr Wichtiges vergessen: Um sprachliche Elemente kombinieren zu können, müssen sie zuerst einmal beherrscht werden. So wenig wert, wie es hier scheint, sind Kenntnisse in diesem Bereich also nicht. Richtig bleibt, daß die immer wieder neue Kombination der Elemente und der situationsangemessene Gebrauch – und nicht Elementen-, Komponentenkenntnisse – die Ziele sind. Die (neue) Frankfurter Aufnahmeprüfung hat deshalb auch nie versucht, nur mit einem Wortschatz- und Grammatiktest zu arbeiten, sondern immer auch auf das Verstehen von Texten und die eigenständige Kombination der Elemente in Texten Wert gelegt.

Während Morrow bloßes Komponentenwissen kritisierte, haben andere die Vorstellung der Mehrdimensionalität der Fremdsprachenfähigkeit gänzlich in Frage gestellt. Bolton nennt Oller als Vertreter der Hypothese einer allgemeinen Kompetenz oder generellen Sprachfähig-

keit, die ein Testen von unterschiedlichen Einzelfertigkeiten überflüssig mache (Bolton 1997: 18). Tests zur Erfassung einer solchen globalen Sprachkompetenz sind nach Darstellung Grotjahns (1997: 121) der C-Test und der Cloze Test. In Frankfurt ist in Vorkursen am Studienkolleg wiederholt mit C-Tests experimentiert worden. Die besten zwei und die schlechtesten zwei von durchschnittlich 15 Teilnehmern konnten dadurch ziemlich verläßlich ermittelt werden. Die Teilnehmer zwischen diesen vier Plätzen aber wiesen in jedem C-Test andere Rangplätze auf. Die Versuche, durch C-Tests die (globale) Sprachkompetenz festzustellen, verliefen also enttäuschend. Wichtiger noch erscheinen aber die folgenden Beobachtungen jedes Praktikers: Gute Grammatikkenntnisse können mit schlechtem Hörverstehen einhergehen, gute mündliche Kommunikationsfähigkeit entspricht oft nicht einer guten Schreibfertigkeit, gutes Leseverstehen bedeutet keineswegs immer auch gutes Hörverstehen usw. Wie plausibel ist dann die Annahme einer generellen Fremdsprachenfähigkeit, die das Erfassen von Teilkompetenzen entbehrlich macht?

Bachman listete 1988 eine Reihe von Untersuchungen auf, die in den achtziger Jahren die auch von ihm selbst vertretene These des »multicomponential view of language ability« (S. 179) unterstützten, und 1998 führte er eine Serie von Tests an, die in den neunziger Jahren auf der Grundlage dieser These entwickelt wurden (S. 7). Brindley erklärt kategorisch:

»Language performance is now recognized by both SLA and LT researchers as being highly complex, multidimensional, and variable according to a variety of social and contextual factors […]« (Brindley 1998: 134)

Die Bemühungen, mit der Aufnahmeprüfung eine ganze Anzahl verschiedener Komponenten zu erfassen, finden also in der Literatur genügend Rückhalt. Dabei heißt »verschiedene Komponenten« immer noch Wortschatz, Grammatik, Landeskunde und die vier Fertigkeiten. Dazu kommen aber u. a. auch kognitive und strategische Komponenten (z. B. bei den unterschiedlichen Höraufgaben), textuelle und pragmatische Komponenten (z. B. beim Verarbeiten von narrativen vs. akademischen Texten).

Heiß umstritten bleibt freilich, welche Komponenten, gegebenenfalls in welcher Kombination und Gewichtung, mit welchen Texten und welchen Aufgaben getestet werden sollen. Bei diesen Fragen allgemeingültige Antworten zu finden ist insbesondere dort ein ernstes Problem, wo ein Test der einzige Kontakt zwischen Tester und Testabnehmer ist. Wer mag da mit letzter Sicherheit entscheiden, daß tatsächlich das Richtige richtig getestet wurde? Wenn, wie das im Studienkolleg der Fall ist, im Unterricht augenfällig wird, ob vorher richtig getestet wurde, erscheinen diese Probleme geringer. Und sollte sich z. B. irgendwann erweisen, daß Friedensengagement oder Kooperationsfähigkeit oder interkulturelle Kompetenz oder ein Transkript, Exzerpt oder Kurzvortrag unverzichtbarer Teil der Prüfung sein muß, steht einer Revision, die aus einer verläßlichen Prüfung eine noch verläßlichere macht, wenig im Weg.

2.3. Bewertung, Erfahrungen

2.3.1 Die Objektivität der Prüfung

a) Da in Frankfurt wegen vergleichsweise großer Bewerberzahlen in mehreren Sälen gleichzeitig geprüft werden muß, wurde auf die *Durchführungsobjektivität* schon früh großer Wert gelegt. Ein Instruktionsblatt hält z. B. genau fest, in welcher Reihenfolge, wie lange, mit welchem Tafelanschrieb und welchen Erläuterungen die Prüfungsteile durchzuführen sind. Es bestehen Vereinbarungen

z. B. über die Sitzordnung, über ein bloßes Minimum an Kommunikation zwischen Prüfenden und Prüflingen während der Prüfung und über die Geschwindigkeit beim Vorlesen der Texte. Außerdem war stets gewährleistet, daß nicht ein Prüfer allein einen Saal betreut.

b) Was die *Auswertungsobjektivität* angeht, so ist diese in den verschiedenen Prüfungsteilen begreiflicherweise unterschiedlich groß. Im Hörtext ist sie vom Prinzip her vollkommen, weil die Zuschreibung und Verrechnung der Punkte den Auswertern keinen Ermessensspielraum läßt. Im Prüfungsteil Wortschatz/Grammatik ist die Auswertungsobjektivität zwar im Laufe der Jahre immer weiter gewachsen, weil immer umfangreichere Lösungsschlüssel den Bewertungsfreiraum der Prüfer begrenzen. Dennoch treten bei jedem Prüfungsdurchgang Lösungsvorschläge auf, die bis dahin nicht angeboten wurden und deren einheitliche Bewertung erst bei der nächsten Durchführung der Prüfung gesichert werden kann. Beim Schreibtest in Form einer Textwiedergabe ist die Objektivität, wie bei diesem freien Testformat nicht anders zu erwarten, am geringsten. Aber auch hier ist die Subjektivität der Bewertung begrenzt. So gibt es eine Anleitung für die Bewertung des Inhalts und eine für die Bewertung der Sprache. Bei der Sprache ist u. a. geregelt, welche Art von Verstoß gar nicht, nur als halber oder aber als ganzer Fehler gewertet wird. Es wird ein Fehlerquotient berechnet, und für bestimmte Fehlerquotienten werden bestimmte Punktzahlen vergeben. Beim Inhalt gibt es eine Liste, die festlegt, welche inhaltlichen Aussagen erwartet werden und wie sie zu bepunkten sind. Im übrigen erfolgt im Anschluß an jede Prüfung vor der Auswertung der Schreibtests eine gemeinsame modellhafte Bewertung mehrerer Arbeiten durch alle an diesem Prüfungsteil beteiligten Auswerter.

2.3.2 Die Reliabilität der Prüfung

Wie steht es mit der Reliabilität, also der »Zuverlässigkeit, mit der bei einer wiederholten Messung unter gleichen Bedingungen dasselbe Meßergebnis herauskommt«? (Rost 1996: 31)

a) Die konstruktbezogene Reliabilität

Bei allen Bemühungen um Objektivität ist dennoch klar, daß das Meßergebnis bei einer Prüfung, die eine Textwiedergabe einschließt, selbst bei ein und demselben Auswerter nie völlig identisch ist. Das führt nicht nur zu einem (begrenzten) Mangel der Auswertungsobjektivität, sondern ist auch ein Mangel der Prüfungskonstruktion. Sie hätte ja auf Sprachproduktion verzichten und sich z. B. auf Multiple-choice-Aufgaben beschränken können. Mit der Entscheidung für freies Schreiben nehmen wir diesen Konstruktmangel in Kauf, der allerdings (siehe oben) in der Praxis beherrschbar, insofern also nicht fundamental zu sein scheint. Für die zwei anderen Prüfungsteile gilt dieser Mangel nicht (Hörtest) bzw. nur bedingt (Lückentexte für Wortschatz/Grammatik); hier ist also vollkommene bzw. weitgehende Konstruktreliabilität gesichert. Zweitens, bei allen drei Prüfungsteilen, also auch beim Schreibtest, wird darüber hinaus auch ein weiteres Kriterium der Konstruktreliabilität beachtet, nämlich die vergleichsweise Länge der Prüfungsteile, denn »es kann prinzipiell davon ausgegangen werden, daß ein Test umso zuverlässiger ist, je mehr Items und Aufgaben er aufweist« (Glaboniat 1998: 35). Drittens ist die Unabhängigkeit vieler Items zu erwähnen. Das gilt z. B. für die Einzelsätze im Strukturtest und für alle Hörtests. Wer bei einer Aufgabe scheitert, erhält mit der nächsten eine echte neue Chance. Ein

zufälliges Scheitern an einer Stelle bedingt also nicht einen (teilweisen/völligen) Punktverlust im weiteren Verlauf des Tests. In abgeschwächter Form gilt das auch für die fortlaufenden Texte in den Strukturtests und ebenso für die Textwiedergabe.

b) Methoden zur Erfassung der erreichten Reliabilität

Die gängigen Methoden zur Erfassung der erreichten Reliabilität, nämlich Textwiederholung (die Prüflinge lösen dieselben Aufgaben ein zweites Mal), Paralleltests (die Prüflinge bearbeiten zwei Tests, von denen der zweite dem ersten in seiner Konstruktion möglichst ähnlich ist), Teiltestvergleich (die Prüflinge bearbeiten einen Test und der Tester vergleicht die Ergebnisse z. B. der ersten mit der zweiten Testhälfte) konnten nicht durchgeführt werden. Das liegt u. a. an der Schwierigkeit, einer identischen Testpopulation denselben Test zweimal anzubieten oder wirkliche Paralleltests zu konstruieren oder bei der Textwiedergabe zu entscheiden, welche Textteile denn zu vergleichen wären. In der »Schule« Studienkolleg fehlten allerdings auch die zeitlichen und finanziellen Voraussetzungen für solche Untersuchungen.

Möglich war es dagegen, die Ergebnisse, die mit ein und demselben Prüfungssatz zu verschiedenen Prüfungsterminen erreicht wurden, zu erfassen und zu vergleichen. Die Prüfung war also identisch, die Testpopulation ähnlich, nämlich Studienbewerber mit nicht weniger als Zertifikatsniveau, die sich um die Aufnahme ins Studienkolleg bemühten, die Prüfer weitgehend identisch und das erfaßte Merkmal gleich: Wieviel Prozent der Prüflinge erreichen 55% der Gesamtpunktzahl? Vier der acht Prüfungssätze konnten an zwei bzw. drei Prüfungsterminen verglichen werden:

Prüfungssatz 2:	36,3%	39,2%	
Prüfungssatz 4:	42,2%	45,2%	41,9%
Prüfungssatz 5:	36,8%	39,7%	
Prüfungssatz 6:	45,2%	45,7%	45,9%

(D. h. also z. B.: Beim Prüfungssatz 2 erreichten einmal 36,3% und ein anderes Mal 39,2% der Prüflinge 55% der Gesamtpunktzahl.)

Der Vergleich der Ergebnisse ergibt, daß im besten Fall, beim Prüfungssatz 6, die Meßergebnisse bei drei Prüfungsterminen erst in der Stelle nach dem Komma voneinander abweichen, und auch im schlechtesten Fall, dem Prüfungssatz 4, beträgt die Abweichung zwischen bestem und schlechtestem Ergebnis lediglich 3,3. Obwohl also die Testpopulation in allen Fällen zwar ähnlich, aber natürlich nicht gleich war, sind »bei einer wiederholten Messung unter gleichen Bedingungen« in einem Fall identische und in den anderen Fällen sehr ähnliche Meßergebnisse herausgekommen. Ich schließe daraus, daß die Reliabilität der Testserien sehr hoch ist.

Natürlich wäre es wünschenswert, wenn diese Annahme in weiteren Durchläufen bestätigt werden könnte. Zur Zeit ist das allerdings nicht möglich, weil sich die Frankfurter Zulassungsstelle dramatisch gestiegenen Bewerberzahlen gegenübersieht und im Gegensatz zu früher weder eine Augenscheinvalidierung behaupteter Deutschkenntnisse vornehmen noch eine Überprüfung der Nachweise über solche Kenntnisse gewährleisten kann. Das bedeutet: Bei einer größeren Zahl echter oder weitgehender Anfänger als Prüfungsteilnehmer sind die Populationen heutiger und früherer Prüfungen nicht mehr vergleichbar und damit Aussagen über Meßergebnisse wertlos.

(Durchschnittliche Teilnehmerzahl bei den letzten 5 Prüfungsterminen: 235 Teil-

nehmer, bei den 5 Terminen davor: 178, bei den 5 Terminen davor: 130, bei den 5 Terminen davor: 171 Teilnehmer. Verglichen mit der Zahl der Teilnehmer ist die Zahl der Bewerber, und deren DaF-Voraussetzungen wären ja zu überprüfen, noch erheblich höher. Sie betrug bei den letzten Prüfungen deutlich mehr als das Doppelte.)

2.3.3 Die Validität der Prüfung

a) Die Augenscheinvalidität (face validity) bezieht sich auf den (ersten) Globaleindruck, den Prüfer, Prüflinge, Experten, Lehrer, Betroffene aller Art von einem Test haben (Glaboniat 1998: 23–25). Ihr Wert ist begrenzt, sie wird gelegentlich auch als ›faith validity‹ ironisiert und kann doch einen ersten Eindruck davon geben, ob eine Prüfung z. B. als angemessen, sinnvoll, plausibel erscheint. Einen Test z. B. mit nichts anderem als 20 Grammatikaufgaben, einen Test, der nur das Hören überprüft, einen C-Test, der aus einem einzigen Text von 60 Wörtern mit 20 Lücken besteht, eine Schreibaufgabe, die in fünf Minuten gelöst sein muß, wird sicherlich kein Experte als valide einstufen, wenn die notwendigen Sprachkenntnisse für die Aufnahme in ein Studienkolleg zu erfassen sind. Nach der obigen Beschreibung, z. B. in der tabellarischen Übersicht, darf die Frankfurter Aufnahmeprüfung wohl mit einer ersten globalen Akzeptanz rechnen.

b) Die inhaltliche Validität

»Inhaltliche Validität ist dann gegeben, wenn ein Text bzw. seine Elemente so beschaffen sind, daß sie […] eine interessierende Verhaltensweise repräsentieren« (Preußler 1997: 13). Sind also Texte und Aufgaben der Frankfurter Prüfung in Art, Inhalt und Umfang so beschaffen, daß sie ein Bild der Fähigkeiten der Prüflinge z. B. im Schreiben, Hören, in der Grammatik und im Wortschatz vermitteln können? Inhaltliche Validität ist nicht errechenbar, sondern muß von Experten, die die von der Testpopulation erwarteten Fähigkeiten kennen, beurteilt werden. So würde ein Experte, der das Abfassen eines formellen Briefes oder die Auswertung und sprachliche Darstellung von statistischem Material als unverzichtbare Fertigkeiten für den Entritt ins Studienkolleg betrachtet, die Frankfurter Prüfung ebenso als nicht valide einstufen wie ein Experte, der in den vorgelegten Texten fachliche Inhalte (der Biologie, Chemie, Soziologie etc.) erwartet.

Experten sollten sich möglichst an eigenen praktischen Erfahrungen und den Erfahrungen von Kollegen aus vergleichbaren Lernsituationen orientieren. Sie werden dabei, zumindest indirekt, immer auch entsprechende Lernzielkataloge berücksichtigen. Brauchbare Orientierungshilfen sind die Rahmenpläne DaF der Lehrgebiete bzw. Studienkollegs, die den Unterricht *nach* der Aufnahmeprüfung bis zur Feststellungsprüfung bzw. DSH beschreiben (Interne Publikationen, Elmau 1977, Regensburg 1979, Bonn 1997). Der Rahmenplan von 1997 gibt darüber hinaus Hinweise zur Aufnahmeprüfung und zu den Vorkursen, an deren Ende die Aufnahmeprüfung steht. Zur Aufnahmeprüfung heißt es u. a. (30): »In der Regel findet eine schriftliche Prüfung statt, die aus mehreren Teilbereichen besteht.« Erwähnt werden die folgenden Bereiche: Wortschatz und Grammatik, Schriftliche Produktion, Leseverstehen, Hörverstehen. Zusätzlich werden die visuellen bzw. sprachlichen Vorgaben für die einzelnen Prüfungsteile und die Aufgabentypen dazu aufgelistet. Die Frankfurter Prüfung stimmt nicht überall mit diesen Hinweisen überein. Insgesamt ist sie, gemessen an diesen Vorgaben, jedoch zweifellos inhaltlich valide. Von den praktizierenden Experten in Frank-

furt spricht ihr ebenfalls keiner die Inhaltsvalidität ab.

c) Die Konstruktvalidität

»Ein Test […] ist konstruktvalide, wenn nachgewiesen werden kann, daß das überprüft wird, was vorgegeben wird, überprüft zu werden« (Glaboniat 1998: 26).

»Dabei spielen neben empirisch-korrelationsstatistischen auch experimentelle Ansätze eine wichtige Rolle. Zur Konstruktvalidierung müssen also vielfältige Untersuchungen durchgeführt werden« (Preußler 1997: 13).

Glaboniat äußert sich zunächst ähnlich, weist aber außerdem darauf hin, daß im Bereich des Sprachentestens mit solchen Verfahren

»oft nur Annäherungswerte geschätzt und errechnet werden, die den großen Aufwand nicht lohnen. Daher verläßt man sich diesbezüglich heute in zunehmenden [sic!] Maße wieder mehr auf ›nicht-rechnerische‹ Validierungsverfahren […]« (1998: 27).

Daß an einer Institution wie dem Studienkolleg die entsprechenden aufwendigen Untersuchungen nicht zu leisten waren, erscheint im Licht der Aussage von Glaboniat nicht als allzu großes Defizit. Das gilt umso mehr, als eine der alternativen Möglichkeiten nicht-rechnerischer, nichtkorrelativer Konstruktvalidierung am Studienkolleg genutzt wurde, nämlich die Gegenüberstellung von Extremgruppen. Hinter diesem Verfahren steht die Überlegung, »daß bei konstruktvaliden Prüfungen die Anfängergruppe markant schlechtere Werte hat als Weit-Fortgeschrittene« (Glaboniat 1998: 28). In Frankfurt wurde einem Fortgeschrittenenkurs wenige Tage vor seiner Abschlußprüfung eine der Aufnahmeprüfungen zur Bearbeitung vorgelegt. Dem Kurs war dieser Prüfungssatz unbekannt. Ergebnis: 50% des Kurses erreichten mindestens 94% der maximal erreichbaren Punktzahl. 80% des Kurses kamen auf mindestens 79,5% der Maximalpunktzahl. Rechnet man die Prüfungsergebnisse aller Kursteilnehmer zusammen, so wurden 86,6% der maximal möglichen Punkte erreicht. Die weiter oben erwähnten Zahlen, wieviel Prozent der »Anfänger« 55 und mehr Prozent der möglichen Punktzahl in vier verschiedenen Frankfurter Prüfungssätzen erreichten, zeigen exemplarisch die markant schlechteren Werte der »Anfängergruppen«.

d) Die kriterienbezogene Validität

Zur Feststellung der kriterienbezogenen Validität wird das Meßergebnis in einem Text in Beziehung gebracht mit einem Außenkriterium. Ein solches Außenkriterium können die Leistungen in einem anderen Test, einer anderen Prüfung, Leistungsnachweise in Klausuren und Beurteilungen von Prüflingen durch Lehrer sein. Natürlich stellt sich hier die Frage, wie valide denn die anderen Tests, Prüfungen, Leistungsnachweise und Beurteilungen ihrerseits sind. Einem einzelnen Außenkriterium, sofern es sich nicht um einen standardisierten Test handelt, wird man sicherlich mit Vorsicht begegnen. Anders ist es, wenn das Ergebnis im Ausgangstest, hier der Aufnahmeprüfung, gleich mit mehreren Außenkriterien (nicht) übereinstimmt. Und so wie der Übereinstimmung mit einem standardisierten Test ein anderes Gewicht zukommt als der Übereinstimmung mit einem informellen Test, so ist auch ein einzelnes Lehrerurteil, das auf einer vergleichsweise kurzen Beobachtung beruht, anders zu gewichten als Urteile mehrerer Lehrer in einem Beobachtungszeitraum von z. B. einem Jahr.

In Frankfurt konnten u. a. die folgenden Außenkriterien herangezogen werden. Erstens, das Lehrerurteil sowohl von DaF- wie auch von Fachlehrern bestätigt, daß sich erfolgreiche Aufnahmeprüflinge im Unterricht sprachlich behaupten kön-

nen. Wenn unterdurchschnittliche Sprachkenntnisse beklagt werden, ließ sich das wiederholt auf die Aufnahmepraxis beim entsprechenden Prüfungstermin zurückführen, sprich: Es wurden Bewerber aufgenommen, deren Punktzahl eine Aufnahme nur bedingt rechtfertigte. Zweitens, schlechte DaF-Noten in der Abschlußprüfung (= Feststellungsprüfung = FSP) stimmten empirisch nachprüfbar mit schlechten Ergebnissen in der Aufnahmeprüfung überein. Ein Nachweis dazu: In meinen letzten zehn Kursen lag der Notendurchschnitt in der FSP-Prüfung DaF bei 2,7. Einer der Kurse erreichte in der FSP lediglich die Durchschnittsnote 3,6. Die Teilnehmer dieses Kurses stammten mit einer Ausnahme von einer sogenannten Warteliste und hatten die reguläre Bestehensgrenze in der Aufnahmeprüfung nicht erreicht. Drittens, gute wie schlechte Aufnahmeprüfungsergebnisse sind generell ein brauchbarer Indikator für den Erfolg in der FSP-Prüfung. Untersucht wurden in diesem Zusammenhang alle FSP-Prüflinge eines Jahrgangs, von denen sowohl das Ergebnis der Aufnahme- wie der FSP-Prüfung vorlag. Bei nur vier (von 65) Kollegiaten ergab sich eine Abweichung bei den beiden Prüfungen um zwei Noten. Bei den anderen stimmten die Noten überein oder sie wichen um nur eine Notenstufe voneinander ab. NB: Abweichung um eine Note kann z. B. heißen, daß die eine Prüfung mit einer 2–, die andere mit einer 3+ abgeschlossen wurde.

2.3.4 Gütekontrolle auf Item-Ebene

Während in den vorangegangenen Abschnitten die Qualität des Gesamttests erörtert wurde, sollen die folgenden Ausführungen der Testgüte auf der Ebene des einzelnen Items gelten. Im Mittelpunkt steht dabei die Frage nach der Itemschwierigkeit sowie nach der Itemtrennschärfe.

Über Jahre hinweg wurden unter dem Gesichtspunkt der angemessenen Schwierigkeit in allen drei Teilbereichen (Struktur-, Hör-, Schreibtest) immer wieder Modifikationen vorgenommen. So wurden z. B. bei Schreibtests einzelne Texte ganz verworfen oder Teile von Texten umgeschrieben oder die Wörterlisten reduziert bzw. erweitert. Ähnlich war es bei den Hörtests, wobei hier zusätzlich einzelne Aufgaben zu den Texten entfernt bzw. verändert wurden. Keine dieser Änderungen beruhte allerdings auf einer systematischen Auswertung und Überarbeitung. Geändert wurde, wo sich Texte oder Aufgaben unübersehbar als zu leicht oder zu schwer erwiesen.

Auch bei den Strukturtests erfolgten immer wieder solche spontanen »subjektiven« Modifikationen. Darüber hinaus waren in diesem Bereich aber auch systematische empirische Untersuchungen möglich. Von den (geschätzt) 300 verschiedenen Struktur-Items der Aufnahmeprüfungen wurden 100 nach den Berechnungshinweisen von Wendeler (1973: 41 f.) hinsichtlich ihrer Schwierigkeit und Trennschärfe gemessen (mehr als 90% der Lösungen richtig = zu leicht; weniger als 30% richtig = zu schwer; Differenz bei der Anzahl richtiger Lösungen zwischen besserer und schlechterer Hälfte der Teilnehmer weniger als 10 % = nicht trennscharf). Die zu leichten, zu schweren, nicht trennscharfen Items wurden entfernt, Ersatz-Items entworfen und beim nächsten Aufnahmeprüfungstermin überprüft und, soweit notwendig, ihrerseits wieder durch neue Items ersetzt. In einer zweiten Untersuchung wurden 100 weitere Struktur-Items erfaßt, bislang jedoch lediglich hinsichtlich ihrer Schwierigkeit (nicht ihrer Trennschärfe) bewertet. Auch Ersatz-Items wurden hier noch nicht entworfen.

3. Zukunft

Das Frankfurter Studienkolleg war bis vor kurzem dem staatlichen Schulwesen zugeordnet und deshalb kaum in der Lage, pädagogische Entwicklungsarbeiten voranzutreiben. Nach der jetzt erfolgten Integration des Kollegs in die Universität könnte sich das ändern, nicht nur im Bereich der Aufnahmeprüfung. Im Bereich dieser Prüfung wäre vor allem die systematische Erfassung aller zu leichten, zu schweren, nicht trennscharfen Items notwendig, und zwar in den Strukturwie in den Hörtests aller acht Prüfungssätze. Im nächsten Schritt müßten diese Items ersetzt und die neuen Items in der Testpraxis auf ihre Güte überprüft werden. Anders bei den Schreibtests; hier kann die Frage der angemessenen Schwierigkeit am einfachsten über die Revision der als Hilfe angebotenen Wörterlisten angegangen werden. Ein zweiter wichtiger Schritt wäre, das Anspruchsniveau der verschiedenen Prüfungssätze weiter anzugleichen. Entscheidend ist und bleibt allerdings die Reliabilität und Validität der einzelnen Serie. Unterschiedliche Schwierigkeitsgrade von Serie zu Serie können erfahrungsgemäß durch variable Bestehensgrenzen aufgefangen werden. Drittens wäre es wünschenswert, weitere Prüfungssätze zu konzipieren, insbesondere für den Fall, daß auch im Bereich der DSH-Kurse auf die Aufnahmeprüfung zurückgegriffen werden soll.

Das Interesse an verläßlichen Aufnahmeprüfungen ist natürlich nicht auf den Standort Frankfurt beschränkt. Das Thema hat innerhalb der Studienkollegs stets eine Rolle gespielt und wird gerade zur Zeit wieder von einer Arbeitsgruppe der nordrhein-westfälischen Kollegs bearbeitet. Außerdem dürfte sich auch der bereits angesprochene DSH-Bereich für eine solche Prüfung interessieren. Schließlich ist an die privaten Anbieter in der Bundesrepublik und alle Anbieter im Ausland zu denken, die mit großer Wahrscheinlichkeit darauf warten, ihre Vorbereitung auf universitäre Sprachkurse bzw. Studienkollegskurse an einer eindeutigen Meßlatte orientieren zu können. Dem einzelnen Studienkolleg und Lehrgebiet wird mit diesem Beitrag eine Anregung gegeben, wie sie ihre Aufnahmeprüfungen für ihre jeweilige Klientel in ihren jeweiligen Lernsituationen objektiv(er), reliabel(er) und valid(er) gestalten können. Damit ist den privaten Anbietern allerdings nur sehr bedingt geholfen. Sie wären auch in Zukunft mit einer Menge sehr verschiedener Prüfungen konfrontiert. Im übrigen blieben die Fragen der Orientierungsnorm, der Vergleichbarkeit und Anerkennung von Prüfungen anderer Institute auch dann weiter offen, wenn die einzelnen Prüfungen hinsichtlich der angesprochenen Qualitätskriterien überprüft und gegebenenfalls verbessert würden. Eine wirkliche Lösung brächte hier nur die Entwicklung einer gemeinsamen Aufnahmeprüfung. Eine Zusammenarbeit von Studienkollegs und Lehrgebieten, FaDaF, dem DAAD, dem Goethe-Institut und dem Volkshochschulverband und u. U. auch dem TestDaF-Institut könnte vielleicht diesen entscheidenden Schritt vorwärts ermöglichen. Da TestDaF sich gegenwärtig auf die Entscheidung beschränkt, ob die für die Aufnahme eines Studiums erforderlichen Sprachkenntnisse vorhanden sind, bliebe wohl von dieser Seite her genügend Raum für Interessenten, die die Sprachkenntnisse auf dem Niveau Ende Grundstufe / Anfang Mittelstufe hinlänglich differenziert und verbindlich erfassen wollen.

Der Versuch, die Frankfurter Aufnahmeprüfung zu überarbeiten und zu ergänzen, kann aus den unterschiedlichsten Gründen scheitern. Der Versuch, eine standardisierte Aufnahmeprüfung für

größere Abnehmergruppen zu schaffen, ist mit Sicherheit noch viel schwieriger. Genug Probleme also und viele Unwägbarkeiten. Immerhin, eins ist sicher: Sollten alle Bemühungen fehlschlagen, dann gibt es in Frankfurt immer noch rund 300 grammatische und lexikalische Items, die – nach Schwierigkeiten gestuft – eine ideale Grundlage für entsprechende Übungsbroschüren wären, Motto: Teste dein Deutsch!

Literatur

Bachman, L. F.; Cohen, A.D. (Hrsg.): *Interfaces between second language acquisition and language testing research*. Cambridge 1998.

Bachman, L. F.: »Language testing – SLA research interfaces«. In: Journal Review of Applied Linguistics 9 (1988), 193–209, abgedruckt in: Bachman; Cohen (Hrsg.) 1988, 177–195.

Bachman, L. F.; Cohen, A. D.: »Language testing – SLA interfaces: An update.« In: Bachman; Cohen (Hrsg.) 1998, 1–31.

Bolton, Sibille: »Tests im Wandel theoretischer Prämissen«. In: Gardenghi; O'Connell (Hrsg.) 1997, 17–23.

Brindley, G.: »Describing language development? Rating scales und SLA«. In: Bachman; Cohen (Hrsg.) 1998, 112–140.

Gardenghi, M.; O'Connell, M. (Hrsg.): *Prüfen, Testen, Bewerten im modernen Fremdsprachenunterricht*. Frankfurt a. M. 1997.

Glaboniat, M.: *Kommunikatives Testen im Bereich Deutsch als Fremdsprache: eine Untersuchung am Beispiel des österreichischen Sprachdiploms*. Innsbruck 1998.

Grotjahn, Rüdiger: »Der C-Test: Neuere Entwicklungen«. In: Gardenghi; O'Connell (Hrsg.) 1997, 117–128.

Lado, Robert: *Testen im Sprachunterricht*. München 1971, übers. R. Freudenstein.

Morrow, K.E.: »Communicative language testing: revolution or evolution?« In: Brumfit, C.J.; Johnson, K.: *The communicative approach to language teaching*. Oxford 1979, 143–157.

Plauen, E.O.: *Vater-und-Sohn-Geschichten*. Konstanz 1949, 4. Aufl. Ravensburg 1964.

Preußler, W.: »Grundbegriffe der klassischen Testtheorie«. In: Gardenghi; O'Connell (Hrsg.) 1997, 11–16.

Rost, J.: *Lehrbuch Testtheorie, Testkonstruktion*. Bern 1996.

Valette, R. M.: *Modern Language Testing*. New York 1967.

Wendeler, U.: *Standardarbeiten – Verfahren zur Objektivierung der Notengebung*. 5. Aufl. Weinheim 1973.

Das Zertifikat Deutsch als Fremdsprache. Hrsg. vom Deutschen Volkshochschul-Verband und Goethe-Institut. Bonn 1972, 3. Aufl. 1985.