Language Testing and Evaluation

18

LTE 18

Rüdiger Grotjahn (Hrsg./ed.)

Der C-Test: Beiträge aus der aktuellen Forschung

The C-Test: Contributions from Current Research

PETER LANG
Internationaler Verlag der Wissenschaften

R. Grotjahn (Hrsg./ed.) · Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research

LANG

Der C-Test: Beiträge aus der aktuellen Forschung
The C-Test: Contributions from Current Research

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn
and Günther Sigott

Volume 18

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Wien

Rüdiger Grotjahn (Hrsg./ed.)

Der C-Test: Beiträge aus der aktuellen Forschung

The C-Test: Contributions from Current Research

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Gedruckt auf alterungsbeständigem, säurefreiem Papier.

ISSN 1612-815X
ISBN 978-3-631-60438-0

www.peterlang.de

Grotjahn, Rüdiger. (Ed.). (2010). Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research. Frankfurt/M.: Lang.

Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung

Thomas Eckes*

The Online Placement Test of German as a Foreign Language (onDaF) is an Internet-delivered gap-filling test based on the C-test principle (www.ondaf.de). The onDaF provides an objective, quick, and reliable measure of general proficiency in German. Its primary uses are to assign L2 learners of German to language courses at institutions of higher education, to provide feedback to L2 learners who plan to take the TestDaF (Test of German as a Foreign Language), and to assist lecturers in deciding on foreign students' eligibility for scholarships of the German Academic Exchange Service (DAAD).

The general design of the onDaF is characterized by the following features: (a) trialling and scaling of a large number of texts (items) by means of a Rasch measurement approach, (b) ongoing construction of a calibrated item bank, (c) placement of test-takers in analogy to the global scale of the Common European Framework of Reference for Languages (CEFR levels A2 to C1), (d) automatic scoring and immediate feedback of test results, (e) worldwide availability through a robust client-server architecture.

In this paper, I elaborate on each of these features. The onDaF is administered at licensed test centers only. This ensures a high level of human supervision and control over the test-taking environment. Each examinee is presented with a unique set of eight texts consisting of 20 gaps each. Texts are drawn from the item bank according to a linear-on-the-fly test delivery model. In each instance, test assembly is subject to the constraints of increasing text difficulty and variation in text topic. The maximum time allowed for each text is 5 min (i.e., maximum test time = 40 min). Responses are automatically scored and test results are reported to examinees immediately after completing the test. A certificate is also issued and stored on the server for later retrieval.

The test construction phase described here comprised a series of 18 separate trialling sessions, covering a total of 3,651 participants from 116 countries. In each session a set of 10 texts was administered. Data collection followed a common-item nonequivalent groups design. Across sets, reliability indices ranged from .94 to .98. Texts showing unsatisfactory model fit or DIF were eliminated. The remaining 135 texts were put on the same difficulty scale through a concurrent estimation procedure. The reliability of the complete set of texts was .97. Based on the proficiency estimates, more than 7 classes of examinees were statistically distinguishable. Cut-scores were set on the onDaF using a specially developed examinee-centered approach, called the prototype group method, combined with a binary logistic regression analysis. Finally, in an external validation study including 223 participants, onDaF placements were compared to placements on the German section of the online language testing system DIALANG. Findings support the claim that the onDaF places learners of German efficiently into four distinct levels of general language proficiency.

* **Korrespondenzadresse:** PD Dr. Thomas Eckes, TestDaF-Institut, Feithstr. 188, D-58084 Hagen. E-mail: thomas.eckes@testdaf.de.

1. Einstufungstest onDaF

Der am TestDaF-Institut entwickelte Online-Einstufungstest Deutsch als Fremdsprache, kurz onDaF, ist ein komplett internetgestützter Lückentest. Sein Aufbau folgt dem C-Test-Prinzip. Ziel ist eine objektive, rasche und zuverlässige Feststellung des Sprachstands von Deutschlernern im In- und Ausland.[1]

Ein wichtiges Einsatzfeld des onDaF ist der universitäre Fremdsprachenunterricht. Ausländische Studierende lassen sich anhand der Testergebnisse nach ihrem allgemeinen Leistungsniveau zu homogenen Lerngruppen bzw. Sprachkursen zusammenfassen. Darüber hinaus dienen die Ergebnisse im onDaF dazu, Lernern eine individuelle Rückmeldung über ihre Sprachkenntnisse zu geben und Lernfortschritte zu dokumentieren. Dies unterstützt die Vorbereitung auf anspruchsvollere und differenziertere Sprachprüfungen wie den TestDaF. Weitere Anwendungsbereiche des onDaF betreffen die Sprachstandsmessung im Rahmen der Prüfung von Bewerbungen um ein DAAD-Stipendium oder auch die Ermittlung der deutschen Sprachkenntnisse im Kontext anderer Testverfahren, wie z.B. der Messung der Studierfähigkeit ausländischer Studienbewerber.

Die Erstellung der im onDaF verwendeten Aufgaben folgt der klassischen Konstruktionsmethode von C-Tests (vgl. z.B. Grotjahn, 2002). In acht kurzen, authentischen Texten werden durch systematische Tilgung von Wortteilen jeweils 20 Lücken erzeugt. Testpersonen haben die Aufgabe, in jedem Text die Lücken korrekt zu ergänzen. Dem so genannten C-Test-Prinzip liegt der Gedanke zugrunde, die Redundanz, die für natürliche Sprachen charakteristisch ist, durch Einfügen von Textlücken zu verringern. Die Leistung, die Testpersonen unter Bedingungen derart reduzierter Redundanz erbringen, erlaubt Aussagen über ihre Kompetenz in der betrachteten Sprache (Eckes & Grotjahn, 2006a; Klein-Braley, 1997; Sigott, 2004).

Sprachkompetenz ist hierbei als eine grundlegende Fähigkeit zu verstehen, die sich aus Wissen und Fertigkeiten zusammensetzt und in vielen verschiedenen Formen des Sprachgebrauchs zum Ausdruck kommt. Die erfolgreiche Bearbeitung des onDaF setzt demnach voraus, dass Testpersonen über ein strukturiertes und differenziertes Sprachwissen verfügen und auf unterschiedliche Komponenten dieses Wissens zugreifen können. Mit anderen Worten, die korrekte Ergänzung der Lücken verlangt die Fähigkeit zur Integration im Gedächtnis gespeicherter Informationen (Top-Down-Verarbeitung) und textspezifischer Informationen (Bottom-Up-Verarbeitung). Eine zentrale Rolle spielen in diesen Sprach-

[1] Aus Gründen der sprachlichen Vereinfachung werden in dieser Arbeit Ausdrücke wie „Lerner“, „Teilnehmer“, „Prüfer“, „Proband“ usw. im generischen Sinne verwendet.

verarbeitungsprozessen Informationen, die sich auf orthografische, lexikalische, morphologische, syntaktische, semantische und kontextuelle Aspekte beziehen.

Damit wird deutlich, was der onDaF messen soll: allgemeine Sprachkompetenz in Deutsch als Fremdsprache. Zugleich wird erkennbar, was der onDaF nicht zu messen beanspruchen kann: Sprachfähigkeit auf der Ebene einzelner Fertigkeiten. Mit anderen Worten, der onDaF erlaubt keine differenzierenden Aussagen über Sprachkenntnisse in den Fertigkeiten des Lesens, Hörens, Schreibens oder Sprechens. Der onDaF ist kein Instrument der Sprachdiagnose, wie sie z.B. Alderson (2005) thematisiert. Das heißt, der onDaF gibt keine Auskunft über Stärken und Schwächen in verschiedenen Komponenten der Sprachkenntnis einer Testperson und auch keine Hinweise auf Sprachbereiche bzw. Sprachfertigkeiten, die gezielt zu fördern oder in ihrer Entwicklung zu beobachten wären. Als allgemeiner fremdsprachlicher Einstufungstest ist dies auch nicht seine Aufgabe.

Für das **generelle Design** des onDaF sind die folgenden Merkmale kennzeichnend: (a) Erprobung und Skalierung einer großen Zahl von Items bzw. Texten auf der Basis des Rasch-Modells (Fischer, 2007; Rasch, 1960/1980), (b) stetiger Aufbau einer kalibrierten Itembank, (c) Einstufung der Sprachkenntnisse analog zur globalen Skala des Gemeinsamen europäischen Referenzrahmens für Sprachen (GER; Europarat, 2001), (d) automatische Testauswertung und sofortige Ergebnisrückmeldung, (e) weltweite und jederzeitige Verfügbarkeit durch eine komplett internetgestützte Testanwendung.

Das zuletzt genannte Merkmal besagt, dass alle relevanten Komponenten des Tests online verfügbar sind. Registrierung der Teilnehmer, Einrichtung und Verwaltung von Testterminen, Buchung von Testterminen, Testbearbeitung und Ergebnisermittlung stützen sich komplett (und ausschließlich) auf das Internet.

Die Online-Komponenten sind zwei separaten Portalen des onDaF zugeordnet: (a) dem Portal für Testabnahmestellen (TAS-Portal), mit den Hauptfunktionen der Termin- und Teilnehmerverwaltung sowie Testdurchführung, und (b) dem Portal für Testteilnehmer (Teilnehmerportal), mit den Hauptfunktionen der Terminauswahl, Testteilnahme und Ergebnisrückmeldung.

Die Auswertung der Teilnehmerantworten ist vollkommen automatisiert, die Rückmeldung der Ergebnisse an die Teilnehmer erfolgt unmittelbar nach Beendigung des Tests. Ein onDaF-Zertifikat, das die Testergebnisse ausweist, steht den Teilnehmern als PDF-Download dauerhaft zur Verfügung.

In der vorliegenden Arbeit gehe ich im Detail auf die zentralen Merkmale des onDaF ein. Nachfolgend behandle ich zunächst die theoretischen und methodischen Grundlagen des onDaF, insbesondere Fragen des Designs internetgestützter Tests, Merkmale und Typen von Itembanken sowie Konzepte und Verfahren

zur Bestimmung von Kompetenzstufen bei C-Tests. Anschließend bespreche ich die Vorgehensweise bei der Erprobung und Skalierung neu erstellter Lückentexte und diskutiere die Ergebnisse einer simultanen Analyse aller erprobten Texte. Zum Schluss stelle ich eine Untersuchung zur Validierung des onDaF dar. Diese Validierungsstudie zielt auf einen Vergleich von Ergebnissen im onDaF mit Ergebnissen im Deutschtest des Online-Testsystems DIALANG (Alderson, 2005; Alderson & Huhta, 2005).

2. Theoretische Grundlagen

2.1. Allgemeine Anforderungen an einen Einstufungstest

Der onDaF zielt darauf ab, Sprachfähigkeit von Testpersonen zu messen und die Testpersonen aufgrund ihrer Ergebnisse im onDaF einer von mehreren Kompetenzstufen zuzuweisen. Um diesem Ziel gerecht zu werden, muss der Test einer Reihe von theoretischen, methodischen und praktischen Anforderungen genügen. Diese Anforderungen leiten sich weitgehend aus den so genannten *Joint Standards* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999) und den darin beschriebenen Testgütekriterien ab (vgl. auch Bühner, 2006; Häcker, Leutner & Amelang, 1998; Kubinger, 1993, 2006; Linn, 2006; Schermelleh-Engel, Kelava & Moosbrugger, 2006).

Ergänzt werden die allgemeinen Gütekriterien durch die von der *International Test Commission* (ITC) ausgearbeiteten Richtlinien für computer- und internetgestütztes Testen (International Test Commission, 2006). Letztere wiederum bauen auf den ITC-Richtlinien für fachgerechte Testanwendung auf (International Test Commission, 2001).

Im Folgenden bespreche ich eine Reihe von Gütekriterien bzw. Anforderungen in ihrem Bezug zum onDaF. Auf die so genannten Hauptgütekriterien, **Objektivität**, **Reliabilität** und **Validität**, werde ich später in den Abschnitten 3 (Konstruktion) und 4 (Validierung) genauer eingehen. Hier sei nur die zentrale Forderung dieser drei Kriterien festgehalten: Die Testergebnisse sollen vom Prüfer oder Untersucher unabhängig sein, nur in geringem Maße Fehlereinflüssen unterliegen und die individuellen Ausprägungen genau desjenigen Merkmals widerspiegeln, das der Test erfassen soll (im vorliegenden Fall: allgemeine Sprachkompetenz).

Ein erstes Kriterium lautet **Skalierung**. Dieses Kriterium ist als erfüllt anzusehen, wenn die Vorschrift zur Verrechnung der Teilnehmerantworten Testwerte liefert, welche die empirischen Merkmalsrelationen adäquat abbilden (auch „Verrechnungsfairness“ genannt; vgl. Kubinger, 1999, 2006). Es ist üblich,

Testwerte auf einfache Weise als Anzahl der richtigen Antworten (Punktzahl, Summenscore) zu bestimmen. Derart ermittelte Testwerte sind aber nur dann verrechnungsfair, wenn das Rasch-Modell gilt. Denn nur bei Gültigkeit des Rasch-Modells bildet der Summenscore eine suffiziente (erschöpfende) Statistik für den Personenparameter, d.h., nur bei Gültigkeit des Rasch-Modells schöpft der Summenscore die gesamte Information aus, die ein Antwortmuster über die Ausprägung der Personenfähigkeit enthält (vgl. ausführlicher hierzu Kubinger, 2006). Alle beim onDaF verwendeten Texte sind nach dem Ratingskalen-Rasch-Modell (Andrich, 1978) skaliert.

Sollen die individuellen Testergebnisse normorientiert, d.h. in Relation zu einer Vergleichs- oder Referenzpopulation ausgewertet und interpretiert werden, dann ist das Kriterium der **Normierung** (auch „Eichung“ genannt) von Bedeutung. Danach müssen die Vergleichswerte (Normen) aktuell sein und die verwendete Stichprobe von Teilnehmern (Eichstichprobe) muss hinreichend groß und repräsentativ für die Referenzpopulation sein. Der onDaF ist aber, wie übrigens auch der TestDaF, ein kriteriumsorientierter Test: Sein Ziel besteht (wie schon gesagt) darin, die Testpersonen nach ihrer Fähigkeitsausprägung einer von mehreren Kompetenzstufen zuzuweisen. Die Frage, die anhand des onDaF beantwortet werden soll, bezieht sich demnach auf das Sprachniveau, das eine Testperson erreicht hat, und nicht auf die Einordnung einer Testperson relativ zur Testwertverteilung einer definierten Population (siehe hierzu Abschnitt 2.3). Bei einer hinreichend großen Zahl von Personen, die den onDaF unter kontrollierten Bedingungen ablegen, lassen sich aber prinzipiell auch für diesen Test Normtabellen konstruieren.

Weitere Anforderungsmerkmale betreffen verschiedene Aspekte der praktischen Anwendung eines Einstufungstests. So sollte der Test in hohem Maße **ökonomisch** sein, d.h., der Test sollte, gemessen an dem zu erwartenden Erkenntnisgewinn, relativ wenige Ressourcen beanspruchen. Hierzu zählen geringe Kosten der Testlogistik (Herstellung und Versand der Testmaterialien, Registrierung von Teilnehmern usw.), hohe Flexibilität (hinsichtlich Ort und Zeit der Testdurchführung), geringe Dauer des Tests selber und rasche Ermittlung bzw. Rückmeldung der Testergebnisse. Der Ökonomieaspekt betrifft auch die Entwicklung von Testaufgaben: Ohne viel Zeitaufwand sollten sich fortlaufend neue Aufgaben erstellen und erproben lassen. Ein internetgestützter C-Test wie der onDaF erfüllt diese spezifischen Anforderungen wie kaum ein anderer Test.

Der Test sollte ferner **nützlich** sein. Damit ist gemeint, dass die Testergebnisse dem Zweck des Tests bestmöglich dienen. Mit anderen Worten, die durch Anwendung des Tests gewonnenen Informationen über die individuelle Fähigkeitsausprägung sollten zu Entscheidungen bzw. Maßnahmen führen, die eine

günstige Nutzen/Kosten-Relation erwarten lassen. Im Falle eines Einstufungstests wie onDaF sollten die Testergebnisse zu Einteilungen von Testpersonen in homogene Lerngruppen führen; zumindest sollten die resultierenden Lerngruppen homogener sein als Gruppen, die nach alternativen Verfahren zustande kämen.[2]

Das nächste Kriterium bezieht sich allein auf die Testteilnehmer: Der Test sollte **zumutbar** sein. Allgemein gesprochen erfüllt ein Test das Kriterium der Zumutbarkeit, wenn er die Testpersonen in zeitlicher, psychischer sowie körperlicher Hinsicht nicht unverhältnismäßig belastet. Es ist einer der wesentlichen Vorzüge computer- oder internetgestützer Tests, dass sie die Testpersonen in aller Regel weniger belasten als herkömmliche Papier-und-Bleistift-Verfahren (Chapelle & Douglas, 2006). Computergestützte Tests ermöglichen beispielsweise eine automatische Auswertung der Teilnehmerantworten und eine sofortige Rückmeldung der Testergebnisse. Sie ersparen damit den Testpersonen lange Wartezeiten und erlauben rasches Handeln im Falle von Sprachdefiziten.

Ein Kriterium, das für Leistungstests, zu denen Sprachtests zu rechnen sind, ein geringeres Problem darstellt als für Persönlichkeits- oder Einstellungsfragebogen, zielt darauf ab, dass Tests **unverfälschbar** sein sollten. Ein unverfälschbarer Test ist so aufgebaut, dass Testpersonen korrekt und wahrheitsgemäß antworten. Verfälschungen sind dann wahrscheinlich, wenn Testpersonen um Selbstauskünfte gebeten werden, das Ziel des Tests durchschauen und motiviert sind, ihre Antworten in eine ganz bestimmte Richtung zu lenken.

Ein Kriterium von genereller Bedeutung besagt, dass ein Test und die im Test enthaltenen Aufgaben **fair** sein sollten. Ein qualitativ hochwertiger Test muss gewährleisten, dass jede Testperson, unabhängig von ihrer Zugehörigkeit zu einer bestimmten Teilpopulation oder Gruppe (z.B. zur Gruppe der Frauen oder Männer oder zu ethnischen Gruppen), die gleiche Chance hat, ein gutes Testergebnis zu erzielen. Ist etwa eine Aufgabe für Frauen schwerer als für Männer, bei gleicher Ausprägung des betreffenden Merkmals bei Frauen und Männern, dann handelt es sich um eine unfaire Aufgabe. In solchen Fällen spricht man auch von einem Item-Bias oder von einer differenziellen Itemfunktion (DIF; vgl. z.B. Clauser & Mazor, 1998; Smith, 2004).

Damit schließt sich die in der pädagogisch-psychologischen Literatur üblicherweise diskutierte Liste von Testgütekriterien. Die Konstruktion des onDaF

---

[2] Nützlichkeit ist hier eines von insgesamt sieben Nebengütekriterien. Dagegen konzipieren Bachman & Palmer (1996) Nützlichkeit (“usefulness”) als zentrales, übergeordnetes Qualitätskriterium, dem sie Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Impact und Praktikabilität subsumieren (vgl. für eine Diskussion dieser Konzeption Alderson & Banerjee, 2002; Grotjahn, 2000).

orientierte sich an diesem Anforderungsprofil. Zusätzlich wurden die besonderen Herausforderungen des internetgestützten Testens berücksichtigt. Im folgenden Abschnitt gehe ich auf einige grundsätzliche Fragen ein, die mit der Nutzung des Internets für Zwecke von Sprachtests verbunden sind, und erläutere die Behandlung dieser Fragen im Kontext des onDaF.

2.2. Testen via Internet

In den letzten Jahren hat das Internet das Design von Tests und die Praxis der Testdurchführung in grundlegender Weise verändert: “The infrastructure is now being built to support a radical change in the way testing is done” (Bartram, 2006b, S. 17). Der Abschnitt, dem das Zitat entnommen ist, trägt die Überschrift “Time for a Revolution!” (Bartram, 2006b, S. 17). Auch wenn man darüber streiten kann, ob die von den enormen Fortschritten der Internet-Technologie ausgelösten Veränderungen im Bereich des Testens „revolutionär“ zu nennen sind, das Internet eröffnet faszinierende neue Wege eines standardisierten und psychometrisch fundierten Testens. Zugleich stellt es aber eine Reihe von beträchtlichen konzeptionellen, methodischen und technischen Herausforderungen (Alderson, 2000; Chapelle & Douglas, 2006; Drasgow & Mattern, 2006; Fulcher, 2003; Jamieson, 2005).

2.2.1. Client-Server-Relation

Wie groß diese Herausforderungen im Einzelfall sind, hängt wesentlich davon ab, wie die Rollen zwischen Client und Server verteilt sind. „Client“ und „Server“ bezeichnen Rechner, die in einem Netzwerk miteinander verbunden sind und unterschiedliche Aufgaben erfüllen. Ganz allgemein lassen sich Clients als Arbeitsplatzrechner oder Arbeitsstationen verstehen, die Daten und Dienste vom Server beziehen.

Im vorliegenden Kontext ist ein **Client** ein lokaler Rechner, an dem Teilnehmer einen internetgestützten Test ablegen. Der **Server** ist der zentrale Rechner, auf dem der Testanbieter die erforderliche Software für die Durchführung des Tests bereitstellt. Je nach Rollenverteilung zwischen Client und Server können die Aufgaben für das Test-Design, aber auch die Chancen, die internetgestütztes Testen für den Entwickler wie für den Anwender bietet, sehr unterschiedlich ausfallen.

So gibt es Tests, bei denen der Server nur die Funktion hat, die für die Testdurchführung notwendigen Instruktionen und Aufgaben zur Verfügung zu stellen. Der Test selber wird aber komplett auf dem Client ausgeführt, einschließlich der Speicherung der Teilnehmerantworten und der Ermittlung der Tester-

gebnisse (so genannter “Fat-Client”). Bei einer anderen Kategorie von Tests übernimmt der Server alle Funktionen der Steuerung des Testablaufs, von der Anmeldung der Teilnehmer über die Auswahl und Darbietung einzelner Aufgaben bis zur Speicherung der Teilnehmerantworten und der Ermittlung und Rückmeldung der Testergebnisse. Der Client-Rechner hat in diesem Fall nur die Aufgabe, Serverdaten auf dem Bildschirm darzustellen und die Eingabe von Daten über Tastatur und Maus zu ermöglichen (“Thin-Client”).

Röver (2001a) spricht von einem Kontinuum der technischen Komplexität (“technological sophistication”) internetgestützten Testens mit “Low-Tech”-Tests, bei denen Clients die Hauptaufgaben erfüllen, am einen Ende und “High-Tech”-Tests, die dem Server alle wichtigen Funktionen zuweisen, am anderen Ende. Erstere sollen im Folgenden auch „clientzentrierte Tests“, letztere „serverzentrierte Tests“ heißen. Der technische Aufwand ist im Falle der serverzentrierten Tests ungleich höher als bei clientzentrierten Tests. Dieser Aufwand ergibt sich insbesondere aus der Notwendigkeit, eine komplexe Client-Server-Architektur aufzubauen und die damit verbundenen Datenbank-Strukturen (für Teilnehmer, Items, Testzentren etc.) zu realisieren. Hierfür ist professionelle Informatik-Unterstützung unerlässlich (vgl. Luecht, 2006). Clientzentrierte Tests können dagegen ohne viel Aufwand mit nur wenigen HTML-Kenntnissen in relativ kurzer Zeit entwickelt werden (vgl. Röver, 2001b, 2002). Im Vergleich zu serverzentrierten Tests bieten sie aber nur ein stark eingeschränktes Spektrum an Funktionen, insbesondere kommen Aspekte der Testkontrolle und Testsicherheit zu kurz oder bleiben gänzlich unberücksichtigt.

Der onDaF ist eindeutig am „High-Tech“-Ende des Kontinuums zu lokalisieren. Die Rollenverteilung zwischen Clients und Server ist stark asymmetrisch. Alle für die Testdurchführung wichtigen Funktionen übernimmt der Server. Die Clients haben nur eine untergeordnete Rolle.

Im Hinblick auf die praktische Testdurchführung sind im onDaF-Netzwerk zwei Arten von Clients zu unterscheiden: Einer der Clients fungiert als Kontrollrechner, die anderen Clients sind die so genannten Teilnehmerrechner. Der Kontrollrechner dient dazu, den gesamten Ablauf einer onDaF-Prüfung zu steuern bzw. zu kontrollieren. Auf dem Bildschirm des Kontrollrechners wird die Liste der zur Prüfung erschienenen und korrekt angemeldeten Teilnehmer angezeigt. Der Testleiter vermerkt auf diesem Rechner das Ergebnis der Identitätskontrolle und schaltet den Test frei. Während der Test läuft, wird der jeweils aktuelle Status der einzelnen Teilnehmer vom Kontrollrechner registriert und angezeigt, d.h., der Testleiter kann zu jedem Zeitpunkt sehen, wer den Test schon beendet hat und wer noch Aufgaben bearbeitet.

Welcher Client als Kontrollrechner dient, ist beliebig. Es ist nur sicherzustellen, dass sich der ausgewählte Rechner im selben Netzwerk wie alle anderen Clients befindet. Diese Voraussetzung wie auch notwendige Einstellungen des Internet-Browsers (z.B. müssen Pop-ups zugelassen werden) werden beim onDaF in einem eigens entwickelten Programm vor dem Start des Tests überprüft. Sobald sich der Testleiter am Kontrollrechner im onDaF-Portal für Testabnahmestellen (TAS-Portal) mit seinen Zugangsdaten angemeldet hat (www.ondaf.de, Link „TAS-Login“), stehen alle Funktionen des onDaF, darunter auch die Kontrollfunktionen für die Testdurchführung, zur Verfügung. Die Teilnehmerrechner dagegen dienen dazu, dass sich die Testpersonen vor Beginn des Tests im onDaF-Portal für Teilnehmer (www.ondaf.de, Link „Teilnehmer-Login“) anmelden und nach Freischaltung des Tests durch den Testleiter die auf dem Bildschirm dargebotenen Texte bearbeiten. Abbildung 1 veranschaulicht den Client-Server-Aufbau, wie er für eine Durchführung des onDaF realisiert ist.

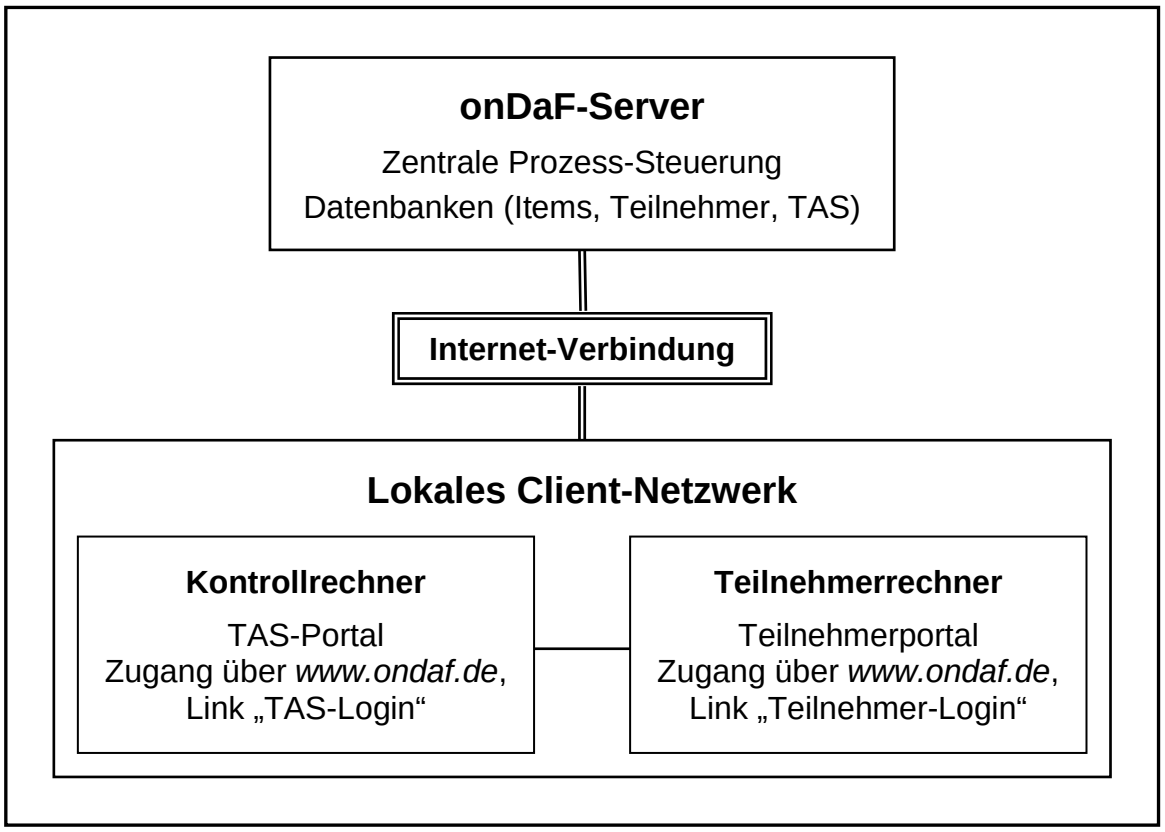

Abbildung 1: Das Client-Server-Netzwerk des onDaF. Kontroll- und Teilnehmerrechner sind Clients, die sich im selben Netzwerk befinden. Der Server verfügt über alle Daten und steuert alle Prozesse, die für eine onDaF-Testung relevant sind. TAS steht für (lizenzierte) Testabnahmestelle.

2.2.2. Testkonsequenzen

Ein anderer Aspekt, an dem sich die Herausforderungen für internetgestütztes Testen bemessen, sind die Konsequenzen der Testergebnisse für Testteilnehmer und andere Personen (z.B. Verwandte, Freunde, Arbeitskollegen). Die Herausforderungen, insbesondere im Hinblick auf die Testsicherheit, sind allgemein umso höher, je schwerer diese Konsequenzen wiegen.

In der Literatur zum Sprachtesten werden traditionell „Low-Stakes-Tests“ und „High-Stakes-Tests“ unterschieden; gelegentlich werden beide Kategorien um die Kategorie der „Medium-Stakes-Tests“ ergänzt (vgl. z.B. Röver, 2001a). Im Falle eines **Low-Stakes-Tests** steht für die Testteilnehmer so gut wie nichts auf dem Spiel. Die Verwendung derartiger Tests bietet sich an für Zwecke der Leistungsrückmeldung, etwa in Form von Self-Assessments, für die Vorbereitung auf andere Tests, die selbst mehr als nur Low-Stakes-Charakter haben, oder auch im Rahmen von Projekten der Sprachtestforschung. Ergebnisse von **Medium-Stakes-Tests** nehmen zwar schon deutlich mehr Einfluss auf die Testteilnehmer, dieser Einfluss beschränkt sich aber auf eher untergeordnete Lebensziele. Typische Beispiele sind Einstufungstests für Fremdsprachler mit anschließender Kursempfehlung oder Abschlusstests im Rahmen der universitären Fremdsprachenausbildung. **High-Stakes-Tests** dagegen beeinflussen übergeordnete Bildungs-, Berufs- und Lebensziele der Testteilnehmer, wie z.B. die Zulassung zu einem Studium, die Einstellung als Mitarbeiter in einem Unternehmen oder die Einbürgerung. Es ist unmittelbar einleuchtend, dass High-Stakes-Tests die mit Abstand höchsten Anforderungen an die Qualität des gesamten Testverfahrens stellen.

Der onDaF ist am ehesten der Kategorie der Medium-Stakes-Tests zuzurechnen. Dies machen die bereits zu Beginn dieses Beitrags aufgeführten Zwecke des Tests deutlich. Als Einstufungstest kann er natürlich auch im Self-Assessment-Kontext Lernern eine wichtige Orientierungshilfe bieten und so Aufgaben eines Low-Stakes-Tests erfüllen. Keinesfalls aber darf der onDaF als High-Stakes-Test verstanden bzw. eingesetzt werden. Dies würde seiner Zielsetzung grundlegend widersprechen.

2.2.3. Testkontrolle

Die Qualität eines internetgestützten Testverfahrens hängt wesentlich von Art und Umfang der Kontrolle ab, die bei der Testanwendung ausgeübt werden kann. Nach Bartram (2006a, 2006b) lassen sich im Hinblick auf die gegebenen Kontrollmöglichkeiten vier Modi der Testdurchführung unterscheiden. Diese Modi haben auch Eingang in die ITC-Richtlinien über computer- und internet-

gestütztes Testen gefunden (Coyne & Bartram, 2006; International Test Commission, 2006). Sie lassen sich wie folgt skizzieren:

(1) **Offener Modus**. Im offenen, unkontrollierten Modus ("open mode") erlauben die Bedingungen weder eine Identifikation der Testteilnehmer noch sehen sie eine Supervision vor, d.h., es gibt keine Testleiter, die Überwachungsaufgaben wahrnehmen könnten. Hierunter fallen Tests, die im Internet frei verfügbar sind und auf die Erfassung von Teilnehmerdaten verzichten. Beispiele sind das Online-Testsystem DIALANG (Alderson, 2005; Alderson & Huhta, 2005; www.dialang.org) und der auf der onDaF-Homepage (www.ondaf.de) angebotene, aus vier Texten bestehende Beispieltest.

(2) **Kontrollierter Modus**. Ähnlich wie im offenen Modus gibt es keine Supervision. Aber der Test ist nur Testteilnehmern zugänglich, die dem Testanbieter bekannt sind ("controlled mode"). Dies wird in der Regel durch Vergabe von Login-Daten (Benutzername, Passwort) an registrierte Teilnehmer gesteuert. Ein Beispiel ist der vom TestDaF-Institut entwickelte C-Test, der im Rahmen des von der RWTH Aachen entworfenen fachbezogenen Self-Assessments eingesetzt wird (www.global-assess.rwth-aachen.de).

(3) **Überwachter Modus**. Dieser Modus ("supervised mode") zeichnet sich dadurch aus, dass es eine Supervision gibt, in deren Verlauf die Testteilnehmer identifiziert und alle Phasen der Testdurchführung kontrolliert werden. Auch unerwartet auftretende Probleme oder Fragen seitens der Testteilnehmer lassen sich im überwachten Modus durch das anwesende Prüfungspersonal behandeln.

(4) **Lizenzierter Modus**. Im lizenzierten Modus ("managed mode") gibt es neben einer strengen Supervision auch eine Kontrolle des institutionellen Kontextes, innerhalb dessen die Testabnahme stattfindet. Diese Form der Kontrolle wird dadurch hergestellt, dass nur jene Institutionen den Test abnehmen dürfen, die zuvor vom Testanbieter lizenziert worden sind. Die Kriterien für eine Lizenzierung als Testabnahmestelle oder Testzentrum betreffen allgemein die räumliche Ausstattung, die technische Ausrüstung oder auch die fachliche Qualifikation des Testpersonals.

Im Falle des onDaF folgt die Testdurchführung dem lizenzierten Modus. Unter der Internet-Adresse http://lizenz.ondaf.de können interessierte Institutionen (universitäre Sprachenzentren, TestDaF-Testzentren, DAAD-Lektorate, Goethe-Institute, Sprachschulen etc.) eine Lizenzierung als onDaF-Testabnahmestelle beantragen. Die im Antrag gemachten Angaben und eventuell den Antrag ergänzende Unterlagen werden vom TestDaF-Institut geprüft. Bei einer positiven Entscheidung wird zwischen dem TestDaF-Institut und der zu lizenzierenden

Einrichtung eine Lizenzvereinbarung getroffen. Hierin werden die Leistungen und Aufgaben des TestDaF-Instituts wie auch die der Einrichtung verbindlich festgeschrieben. Auf Seiten des Antragstellers stehen Aspekte der Kontrolle und der Sicherheit der Testdurchführung im Mittelpunkt. Nach erfolgter Lizenzierung erhält die Einrichtung die Zugangsdaten für die Nutzung des TAS-Portals.

Die wichtigsten Funktionen des TAS-Portals sind: (a) **Terminverwaltung**. In einem Online-Kalender können Prüfungstermine neu angelegt oder bereits angelegte Termine geändert werden. Auch lässt sich die maximal mögliche Anzahl von Teilnehmern angeben und festlegen, bis wann Termine gebucht sein müssen. (b) **TAN-Verwaltung**. Die Buchung eines Termins setzt voraus, dass Teilnehmer eine Transaktionsnummer (TAN) eingeben, die sie von einer TAS ihrer Wahl erhalten haben. In der TAN-Verwaltung können TAS-Leiter eine bestimmte Anzahl von TANs beim TestDaF-Institut anfordern und Listen gültiger TANs ausgeben lassen. (c) **Teilnehmerverwaltung**. Es kann eine Liste der Teilnehmer, die einen bestimmten Prüfungstermin gebucht haben, erstellt werden. Zudem lassen sich die Kontaktdaten der betreffenden Teilnehmer einsehen. (d) **Testdurchführung**. Alle Teilnehmer werden, sobald sie sich vor Beginn der Prüfung an ihren Rechnern eingeloggt haben, auf einem Kontrollmonitor aufgelistet. In dieser Liste wird das Ergebnis der Identitätskontrolle notiert. Vom Kontrollrechner aus wird der Test gestartet und die komplette Testabnahme fortlaufend überwacht. (e) **Ergebnisausgabe**. Nach Abschluss eines Testdurchgangs steht online eine Liste mit den Testergebnissen der Teilnehmer zur Verfügung. In dieser Liste finden sich auch die onDaF-Zertifikate der Teilnehmer in Form von PDF-Dateien.

Im TAS-Portal ist ein Handbuch mit einer ausführlichen Beschreibung der vielfältigen Funktionen des onDaF abrufbar. Das onDaF-Handbuch enthält ferner eine skriptähnliche Anleitung, die auf einer einzigen Seite durch den gesamten Testablauf führt.

2.2.4. Testsicherheit

Ein besonders kritisches Element des internetgestützten Testens betrifft die Frage der Testsicherheit. Schon bei Medium-Stakes-Tests gilt es, alles zu unternehmen, um größtmögliche Sicherheit des Testverfahrens zu gewährleisten. Dieses Ziel lässt sich im Falle von serverzentrierten Tests am ehesten realisieren. Da der onDaF genau dieser Kategorie von Tests angehört, beziehen sich die nachfolgenden Ausführungen ausschließlich auf Tests mit zentraler Server-Steuerung.

Drei Gegenstandsbereiche der Testsicherheit sind zu unterscheiden: (a) das Testmaterial, (b) die Teilnehmeridentität und (c) die Teilnehmerdaten. Abgese-

hen von der Browser-Software (onDaF ist optimiert für Microsoft Internet Explorer und Mozilla Firefox) befinden sich alle für die Testdurchführung relevanten Informationen, d.h. die gesamte Applikationssoftware und alle Daten, auf dem Server und nicht auf den Clients. Damit hat der Testanbieter volle Kontrolle über die Testinhalte (Aufgaben, Instruktionen etc.), das Verfahren zur Auswertung von Teilnehmerantworten, die Ermittlung und Rückmeldung der Testergebnisse usw. Außerdem kann sich der Testanbieter zu jedem Zeitpunkt leicht einen Überblick darüber verschaffen, wer wann welche Testversion verwendet, sofern unterschiedliche Testversionen auf dem Server hinterlegt sind. Da die Testsoftware und wichtige Referenzdaten (beim onDaF sind dies vor allem die Themenkategorien, denen die Texte zugeordnet sind, und die Textschwierigkeiten) nur an einem einzigen Ort existieren, haben alle Testanwender jederzeit Zugriff auf dieselbe Testversion, die zugleich die aktuellste ist. Schließlich lassen sich auch notwendige Veränderungen am Test, Verbesserungen der Benutzeroberfläche, Fehlerkorrekturen usw. auf dem Server relativ einfach vornehmen.

Hinsichtlich des Aspekts der Teilnehmeridentität ist der jeweils praktizierte Modus der Testdurchführung von Bedeutung. Nur im überwachten und im lizenzierten Modus ist eine Identifizierung der Teilnehmer ausdrücklich vorgeschrieben. Diese beiden Modi sind in der Regel mit einem serverzentrierten Testverfahren verknüpft. Im Falle einer onDaF-Testung melden sich die Teilnehmer in der TAS an ihren Rechnern mit Benutzernamen und Passwort an. Alle korrekt eingeloggten Teilnehmer werden auf einem Kontrollmonitor angezeigt. Bevor der Prüfer den Test starten kann, muss er jeden einzelnen Teilnehmer anhand eines Ausweises mit Lichtbild zweifelsfrei identifizieren. Die so identifizierten Teilnehmer werden in der Teilnehmerliste markiert. Erst danach erfolgt die Freischaltung des Tests. Auch wenn der Test selber vollkommen automatisiert abläuft, bleibt der Prüfer anwesend, um den Testverlauf zu überwachen.

Alle Daten, die von den Teilnehmern erzeugt werden, sei es bei der ersten Registrierung, bei der Buchung von Prüfungsterminen oder bei der Bearbeitung der Testaufgaben, werden auf dem Server gespeichert. Auf diese Daten, insbesondere auf die Testergebnisse, haben ausschließlich jene Personen Zugriff, die hierfür eine Berechtigung erhalten haben. Teilnehmer können mit ihren Login-Daten auf die Ergebnisse, die sie im onDaF erzielt haben, zugreifen. Diese Ergebnisse sind in Form eines onDaF-Zertifikats zugänglich. Das Zertifikat verzeichnet den Teilnehmernamen, Geburtsdatum, Datum und Ort der Prüfung, die erreichte Punktzahl und die onDaF-Einstufung.

Da das onDaF-Zertifikat im Teilnehmerportal als PDF-Download zur Verfügung steht und beliebig oft reproduziert werden kann, sollte die Echtheit des Zertifikats durch Dritte, insbesondere durch den DAAD oder Sprachenzentren deutscher Hochschulen, auf einfache Weise überprüfbar sein. Dies leistet ein speziell für den onDaF entwickeltes Online-Verifizierungsmodul (erreichbar unter www.ondaf.de/check). Testabnahmestellen können ihrerseits im TAS-Portal jederzeit die komplette Liste der Teilnehmer mit ihren Testergebnissen einsehen (im HTML-Format), ausdrucken oder weiterbearbeiten (im Excel-Format).

2.3. Konstruktion einer Itembank

2.3.1. Merkmale von Itembanken

Ganz allgemein gesprochen ist eine **Itembank** (gelegentlich auch Item-Pool genannt) eine strukturierte Menge von Items zur Erstellung von Tests. Diese Definition ist so weit gefasst, dass sich ihr verschiedene Formen und Zwecke von Itembanken subsumieren lassen.

Es finden sich in der Literatur häufig engere Definitionen, die das eine oder andere Merkmal von Itembanken in den Mittelpunkt rücken. Beispielsweise betonen Wright & Stone (1999, S. 107) den Aspekt der Messung einer bestimmten Variablen: “An item bank is a set of carefully composed and jointly calibrated items that develop, define and quantify a single common theme and hence provide an operational definition of one variable”. Wolfe (2000, S. 411) rückt die Herstellung äquivalenter Testformen in den Vordergrund: “a set of items from which test forms that create equivalent measures may be constructed”. Die Definition von Ariel, van der Linden & Veldkamp (2006, S. 85) verweist auf den Aspekt der Computerisierung: “a collection of test items for a given domain usually stored in computer memory along with a list of codes for their attributes”.

Die Menge der in einer Itembank abgelegten Items heißt **strukturiert**, weil jedes Item mit bestimmten Merkmalen oder Attributen in systematischer Weise verknüpft ist. Van der Linden (2005) unterscheidet drei Klassen von Item-Attributen: (a) kategoriale Attribute (z.B. Themenkategorie, Antwortformat, Schwierigkeitsstufe), (b) quantitative Attribute (z.B. Schwierigkeitsparameter, Messgenauigkeit, Anzahl der Wörter), (c) logische Attribute (z.B. Beziehungen zwischen Items wie Exklusion oder Inklusion).

Erst eine Itembank mit wohl durchdachter Struktur bietet **Vorteile** wie die folgenden (vgl. Henning, 1987; Lee, 2006; Umar, 1999): (a) Effizienz (leichter Zugriff auf Items sowie unkomplizierte Zusammenstellung von Items zu Tests mit bestimmten Eigenschaften), (b) Standardisierung (sorgfältige, nach definier-

ten Regeln bzw. Kriterien erfolgende Erstellung, Überarbeitung und Erprobung von Items), (c) Ökonomie (insbesondere bei computergestützten Itembanken Verringerung der Kosten ständig neu zu konstruierender Tests), (d) Konsistenz (bei Anwendung von Item-Response-Modellen Erstellung multipler äquivalenter Testformen durch Item-Kalibrierung), (e) Flexibilität (Erstellung von Tests, die hinsichtlich Testlänge, Testinhalt und Schwierigkeitsniveau für die jeweilige Anwendung maßgeschneidert sind).

Um diese vorteilhaften Eigenschaften nutzen zu können, ist allerdings in der Planungs- und Aufbauphase, insbesondere im Hinblick auf die Erstellung, Erprobung und Evaluation einer großen Zahl von Items, ein höherer Aufwand als bei konventionellen Testkonstruktionen erforderlich. Hinzu kommen Kosten für die fortlaufende Aktualisierung, Pflege und Weiterentwicklung der Itembank (Lee, 2006; Szabó, 2008). Dies setzt IT-Kompetenz und Kompetenz in der Anwendung psychometrischer Modelle und Methoden voraus.

2.3.2. Typen von Itembanken

Itembanken lassen sich einmal danach unterscheiden, wie differenziert die psychometrische Methodik ist, die ihrem Aufbau zugrunde liegt. Eine andere Dimension der Unterscheidung betrifft das Ausmaß, in dem Itembanken computergestützt entwickelt und genutzt werden (vgl. Umar, 1999).

Im Hinblick auf die Dimension der methodischen Differenzierung finden sich auf der untersten Stufe Itembanken, deren Items zwar sorgfältig erstellt und von Experten als inhaltlich valide eingestuft werden, die aber darüber hinaus keinerlei psychometrische Analyse erfahren haben. Auf der nächst höheren Stufe sind Itembanken einzuordnen, deren Items empirisch erprobt und nach klassischen Qualitätsmerkmalen (Lösungsrate, Trennschärfe, Qualität der Distraktoren) evaluiert werden. Auf der höchsten Stufe psychometrischer Differenzierung liegen kalibrierte Itembanken. Eine **kalibrierte Itembank** zeichnet sich dadurch aus, dass die in ihr abgelegten Items nicht nur nach klar definierten Kriterien erstellt, inhaltlich valide und empirisch erprobt sind, sondern auch eine hinreichend große Anpassung an ein Item-Response-Modell besitzen (Henning, 1987; Kolen & Brennan, 2004; Lord, 1980). Ist Modellanpassung im Sinne der Item-Response-Theorie gegeben, so können die Parameterschätzungen für alle Items in der Bank auf ein und derselben Dimension ihrer Schwierigkeit angeordnet werden (Methode der Item-Kalibrierung). Dieser Typ von Itembanken bietet ein Maximum an Effizienz, Konsistenz und Flexibilität bei der Konstruktion und Anwendung von Tests (vgl. auch Traxel & Dresemann, im vorliegenden Band).

Nach dem Grad der Computerisierung lassen sich wiederum drei Typen von Itembanken unterscheiden. Auf der untersten Stufe dieser Dimension finden sich

manuelle Itembanken. Der Einsatz von Computern ist in keiner Phase des Prozesses vorgesehen. Stattdessen übernimmt ein mehr oder weniger komplexes System von Karteikarten die Aufgabe, Items z.B. nach Antwortformat oder Inhalt zu sortieren. Die nächst höhere Stufe umfasst teilmanuelle Itembanken. Diese stützen sich zwar immer noch auf ein System von Itemkarten, verwenden aber Computerprogramme zur Bestimmung der Itemqualität. Auch die Organisation der Itembank orientiert sich an den Ergebnissen der Itemanalysen. Auf der höchsten Stufe liegen **computerisierte Itembanken** (Schmeiser & Welch, 2006; Vale, 2006). Die Anwendung von Computerprogrammen beschränkt sich bei diesem Typ nicht auf die Itemanalyse. Vielmehr kommen spezielle Programme zum Einsatz, die dazu dienen, Itembanken flexibel und benutzerfreundlich zu strukturieren, Items zu verwalten (z.B. neue Items aufzunehmen, bereits gespeicherte Items zu revidieren), einzelne Items zu Testformen zusammenzustellen, Antworten von Testteilnehmern automatisch auszuwerten usw. Es versteht sich von selbst, dass die Konstruktion einer kalibrierten Itembank einen sehr hohen Grad an Computerisierung voraussetzt. Vale (2006) listet Computerprogramme für Itembanken mit unterschiedlichen Verwendungszwecken auf. Solche Programme können aber auch bei entsprechender IT-Kompetenz vom Testanbieter selber maßgeschneidert entwickelt werden. Letzteres war bei der Entwicklung des onDaF der Fall.

2.3.3. Aufbau einer kalibrierten Itembank

Wie bereits erwähnt bieten kalibrierte Itembanken ein Höchstmaß an Funktionalität und Effizienz. Um eine solche Itembank zu konstruieren, sind aus psychometrischer Sicht zunächst drei Fragen zu klären.

Die erste Frage betrifft das **Item-Response-Modell**, anhand dessen die Kalibrierung der Items erfolgen soll. Von den vielen in der Literatur diskutierten Modellen (vgl. z.B. Embretson & Reise, 2000; Rost, 2004; Yen & Fitzpatrick, 2006) weist das Rasch-Modell Eigenschaften auf, die es (auch) für den Zweck des Item-Bankings besonders geeignet erscheinen lassen (vgl. Henning, 1987; Umar, 1999; Wolfe, 2000). Zu diesen Eigenschaften zählen (a) die wechselseitige Unabhängigkeit der Schätzungen von Item- und Personenparametern, (b) der schon im Zusammenhang mit dem Kriterium der Skalierung angesprochene Sachverhalt, dass Summenscores suffiziente Statistiken sind, und (c) die relativ geringen Anforderungen an den Umfang der Personenstichprobe bei der Skalierung der Items. Diese und andere vorteilhafte Eigenschaften des Rasch-Modells lassen sich auf das besondere Antwortformat von Tests, die nach dem C-Test-Prinzip konstruiert sind, übertragen (Eckes, 2006a, 2006b, 2007; siehe auch Abschnitt 3.3).

Die zweite Frage bezieht sich auf das Design der Datenerhebung. Das heißt, es ist ein Versuchsplan zu wählen, auf dessen Grundlage die für eine gemeinsame Kalibrierung von Items aus verschiedenen Testformen erforderlichen Daten gewonnen werden. Für einen derartigen Versuchsplan sind Ausdrücke wie **Linking-** oder **Anchoring-Design** gebräuchlich (vgl. z.B. Henning, 1987; Holland & Dorans, 2006; Vale, 1986). Da die anhand eines solchen Plans gewonnenen Daten ebenso gut einer Score-Adjustierung unterschiedlicher Formen desselben Tests, d.h. einem Test-Equating, dienen können, sprechen einige Autoren auch von einem Equating-Design (vgl. z.B. Cook & Eignor, 1991; Kolen & Brennan, 2004; von Davier, Holland & Thayer, 2004; Wolfe, 2000). Zwei prominente Linking-Designs seien kurz vorgestellt.[3]

Im ersten Design, dem **Plan äquivalenter Gruppen** ("Random Groups Design", "Equivalent Groups Design"), bearbeiten wenigstens zwei Gruppen von Testpersonen je eine Testform. Weil zwischen den Personengruppen bzw. zwischen den Testformen keine unmittelbare Verbindung besteht, ist eine gemeinsame Skala von Itemparametern nur konstruierbar, wenn entweder die Personengruppen oder die Testformen als äquivalent betrachtet werden können. Beim Plan äquivalenter Gruppen wird direkte Vergleichbarkeit zwischen den Personengruppen in der Weise hergestellt, dass die verschiedenen Testformen nach Zufall den Testpersonen zugeteilt werden. Abbildung 2 (oberer Teil) illustriert dies am Beispiel von zwei Testformen A und B.

Der Plan äquivalenter Gruppen bietet den Vorteil, dass jede Testperson eine einzige Testform und nur diese zu bearbeiten hat. Allerdings müssen alle Testformen zum Zeitpunkt der Datenerhebung verfügbar sein und zur selben Zeit an die Testpersonen ausgegeben werden. Üblicherweise geschieht die Ausgabe in alternierender Weise, d.h., die erste Testperson erhält Form A, die zweite Testperson Form B, die dritte Testperson wieder Form A usw. (so genanntes "Spiraling"). Darüber hinaus muss die Gesamtzahl der Testpersonen hinreichend groß sein (die erforderliche Anzahl der Testpersonen ist bei diesem Plan deutlich größer als bei allen anderen Plänen). Für Zwecke eines Test-Equatings mögen diese Bedingungen noch unproblematisch sein, im Hinblick auf die Konstruktion einer umfangreichen Itembank erscheinen sie aber nur schwer realisierbar.

---

[3] Kolen & Brennan (2004) verwenden den Ausdruck "Linking" in einer etwas anderen Bedeutung. Sie verstehen darunter ein statistisches Verfahren zur Score-Adjustierung bei Tests, die sich in planvoller Weise nach Inhalt und/oder Schwierigkeit unterscheiden. Den Ausdruck "Equating" beziehen die Autoren auf die Score-Adjustierung bei Tests, die so konstruiert werden, dass sie sich in ihren inhaltlichen und statistischen Eigenschaften möglichst ähnlich sind.

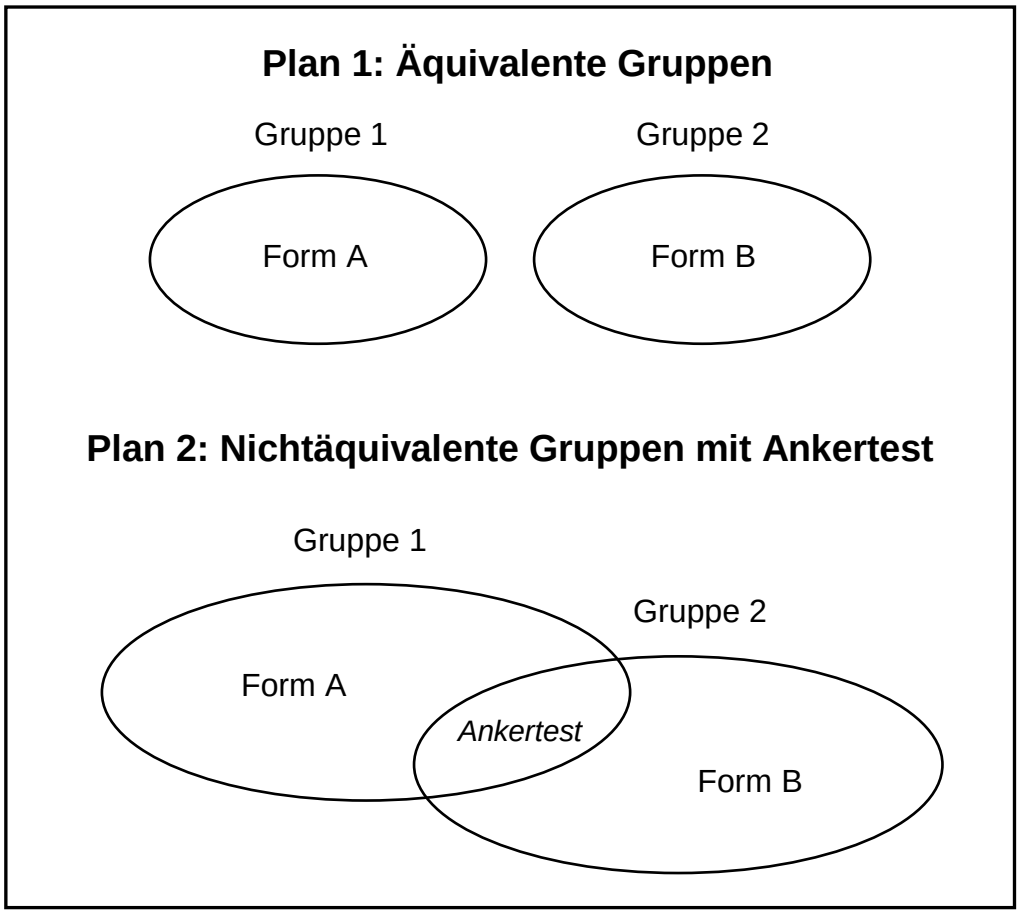

Abbildung 2: Illustration zweier Pläne für die Datenerhebung im Rahmen der Kalibrierung von Items aus verschiedenen Testformen auf einer gemeinsamen Skala. Dargestellt ist der Fall zweier Testformen A und B, die von zwei Gruppen von Testpersonen zu bearbeiten sind. Plan 1 zielt auf Herstellung von Äquivalenz der beiden Gruppen durch zufällige Zuteilung der Formen A und B zu Testpersonen. Bei Plan 2 liegen nichtäquivalente Gruppen vor, d.h. die Gruppen können ein unterschiedliches Fähigkeitsniveau haben (angedeutet durch die Höhendifferenz zwischen den Ellipsen für Formen A und B). Eine gemeinsame Kalibrierung der Items von Formen A und B erfolgt nach Plan 2 auf der Basis des Ankertests.

Ein Design, das sich für gemeinsame Item-Skalierungen im Rahmen des Item-Bankings als sehr nützlich erwiesen hat, verwendet potenziell nichtäquivalente Gruppen von Testpersonen. Die Gruppen können hierbei ein unterschiedliches Fähigkeitsniveau haben. Genauer gesagt, sie können sich hinsichtlich des Mittelwerts und der Standardabweichung der Verteilungen der latenten Variablen voneinander unterscheiden. Um bei gegebenen Gruppenunterschieden eine gemeinsame Skala der Itemparameter zu konstruieren, wird ein Satz von Items, die allen Testformen gemeinsam sind, eingeführt. Dieses Vorgehen charakteri-

siert den so genannten **Plan nichtäquivalenter Gruppen mit Ankertest**, kurz **Ankertestplan** (in der englischsprachigen Literatur auch "Common-Item Non-equivalent Groups Design", kurz CING-Design, oder "Non-Equivalent groups with Anchor Test Design", kurz NEAT-Design). Abbildung 2 (unterer Teil) veranschaulicht die Grundstruktur des Ankertestplans.

Allgemein gilt für dieses Design, dass die Verbindung zwischen den Testformen umso besser ist, je ähnlicher sich die Personengruppen hinsichtlich der Merkmale sind, die das Ergebnis in den Testformen beeinflussen können (z.B. Lernstand, Motivation, demografische Merkmale usw.). Außerdem ist zu beachten, dass die im Ankertest zusammengefassten gemeinsamen Items das gleiche Konstrukt erfassen sollten wie die Items der Testformen.

Die häufig vertretene Auffassung, dass der Ankertest nicht nur gleiche inhaltliche, sondern auch ähnliche statistische Eigenschaften wie die Testformen besitzen sollte (vor allem im Hinblick auf die Verteilung der Itemschwierigkeiten), wird von neueren Simulationsstudien nicht gestützt (Sinharay & Holland, 2006a, 2006b). Danach ist der traditionell empfohlene „Minitest" keineswegs einem Ankertest mit stark eingeschränkter Streuung der Itemschwierigkeiten überlegen.

Was die Anzahl der gemeinsamen Items betrifft, so gehen die Angaben in der Literatur weit auseinander. Die empfohlenen Werte reichen von etwa 20 Items oder 20% der Anzahl von Items einer Testform (Angoff, 1971) bis hin zu nur 2 Items (Vale, 1986; Wingersky & Lord, 1984). Dabei ist zu berücksichtigen, dass die geeignete Anzahl gemeinsamer Items auch von der Länge der Testformen und vom Format des Ankertests abhängt.

Es gibt zwei Varianten des Ankertestplans. In der ersten Variante tragen die Antworten der Testpersonen auf die gemeinsamen Items zum Score im gesamten Test bei; der Satz an gemeinsamen Items wird **interner Ankertest** genannt. In der zweiten Variante ist dies nicht der Fall, d.h., die Antworten auf die gemeinsamen Items werden getrennt verrechnet; dadurch ist ein **externer Ankertest** definiert (vgl. Kolen & Brennan, 2004; Lord, 1980; Wolfe, 2000). Abbildung 2 (unterer Teil) veranschaulicht die erste Variante. In der Regel werden die Items eines internen Ankertests unter die anderen Items einer Testform gemischt. Ein externer Ankertest ist üblicherweise ein separat dargebotener Test. Die Items eines solchen Ankertests können ein Format besitzen, das sich vom Format der Items der Testformen deutlich unterscheidet.[4]

[4] Ein externer Ankertest, ein C-Test, kommt z.B. beim TestDaF zum Einsatz (vgl. Eckes & Grotjahn, 2006b).

Angesichts des beim onDaF verfolgten Ziels, d.h. Aufbau einer kalibrierten Itembank mit einer stetig wachsenden Zahl an Items bzw. C-Test-Texten und zeitlich ausgedehnten Erprobungen von Testformen im In- und Ausland, kam ein Plan mit äquivalenten Gruppen nicht in Betracht. Vielmehr wurde ein Ankertestplan mit einem internen Anker zugrunde gelegt. Abschnitt 3.3 behandelt das konkrete Vorgehen näher.

Schließlich stellt sich neben der Wahl des Item-Response-Modells und des Designs der Datenerhebung die Frage nach der statistischen Methode, mittels derer die Items aus verschiedenen Testformen auf einer gemeinsamen Skala kalibriert werden. Es lassen sich grob zwei Klassen von Methoden unterscheiden: simultane und separate Schätzmethoden (vgl. Kim & Cohen, 1998; Kolen & Brennan, 2004; Lee, Song & Kim, 2004).

Bei **simultanen** Methoden werden die Parameter der Items aus den betreffenden Testformen in einer einzigen, gemeinsamen Analyse geschätzt. Dies erfordert Computerprogramme, die eine Schätzung von Parametern bei multiplen Gruppen mit unterschiedlichen Mittelwerten und Standardabweichungen auf einer gemeinsamen latenten Dimension erlauben. Im Falle der **separaten** Methoden werden pro Gruppe von Testpersonen (bzw. pro Testform) zunächst unabhängige Analysen durchgeführt. Anhand der Parameterschätzungen für die gemeinsamen Items wird eine lineare Skalentransformation ermittelt, die es erlaubt, die Itemparameter einer gegebenen Testform in die Skala der Itemparameter der anderen Testform zu überführen.

Vergleiche zwischen Methoden beider Klassen haben gezeigt, dass die Parameterschätzungen nach simultanen Methoden in aller Regel genauer sind als die nach separaten Methoden (vgl. Hanson & Béguin, 2002; Lee, Song & Kim, 2004). Die Überlegenheit der simultanen Methoden geht vermutlich darauf zurück, dass die Schätzungen der Parameter für die gemeinsamen Items (d.h. für die Ankeritems) auf einer relativ großen Stichprobe basieren. Beim onDaF kam denn auch eine simultane Schätzmethode zum Einsatz.

2.4. Festlegung von Cut-Scores

Die Ergebnisse, die Testpersonen im onDaF erreichen, werden zunächst in Form von Summenscores ausgedrückt. Das heißt, das auf dem onDaF-Server implementierte Programm für die Testauswertung vergibt für jede korrekt ergänzte Lücke genau einen Punkt. Die Summe der Punkte über alle Lücken und Texte bildet den Summenscore einer Testperson. Da der onDaF aus acht Texten mit jeweils 20 Lücken besteht, kann der Summenscore Werte zwischen 0 und 160 annehmen. Gemäß dem C-Test-Konzept indizieren höhere Werte ein höheres Maß an Sprachfähigkeit.

Wie aber sind die Scores, die Testpersonen im onDaF erreicht haben, zu interpretieren? Was bedeutet z.B. ein Score von 98 Punkten? Wie hoch ist die Sprachfähigkeit einer Testperson mit diesem Score einzuschätzen? Erlaubt dieses Ergebnis eine Zuweisung zu einem Sprachkurs für fortgeschrittene Lerner oder ist doch eher die Zuweisung zu einem Mittelstufenkurs die richtige Entscheidung? Bei welchem Score wäre die Grenze zwischen Mittelstufe und Oberstufe anzusetzen?

Fragen dieser Art lassen sich mit Methoden des **Standard-Settings** beantworten. Standard-Setting meint die Festlegung von Testscores, anhand derer Testpersonen in zwei oder mehr Leistungskategorien eingeteilt werden können. Diese kritischen Testscores werden auch als „Cut-off-Scores“ oder kurz „Cut-Scores“ bezeichnet. Im einfachsten Fall wird zwischen den beiden Kategorien „bestanden“ (eine Testperson hat den Cut-Score erreicht oder überschritten) und „nicht bestanden“ (eine Testperson hat den Cut-Score nicht erreicht) unterschieden. Beim onDaF lautet das Ziel, Testpersonen einem von vier Kompetenzniveaus zuzuordnen. Diese Kompetenzniveaus sollen sich darüber hinaus so eng wie möglich an den GER-Niveaustufen orientieren.

Im Folgenden gehe ich auf einige Grundfragen und Methoden des Standard-Settings ein und diskutiere Probleme und Möglichkeiten der Unterscheidung von Kompetenzniveaus im Falle von C-Tests. Anschließend skizziere ich die Vorgehensweise, die beim onDaF Anwendung findet.

2.4.1. Standard-Setting-Methoden

Die Liste der in der Literatur diskutierten Methoden des Standard-Settings ist lang und sie wird von Jahr zu Jahr länger (vgl. z.B. Cizek, 2006; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Zieky, 2001). Gemeinsam ist diesen Methoden, dass sie Urteile oder Einschätzungen von speziell geschulten Beurteilern, Fachkräften oder Experten verlangen. Ziel ist es, diese subjektiven Urteile in Empfehlungen zur Festlegung von Cut-Scores zu bündeln. Dabei ist ein Vorgehen zu wählen, das standardisiert ist, klar definierten Regeln folgt und sich in allen relevanten Phasen an allgemein anerkannten, wissenschaftlichen Kriterien orientiert.

Unterschiede zwischen den Methoden bestehen hauptsächlich in drei Bereichen. Der erste Bereich bezieht sich auf Merkmale der Urteilsaufgabe. Hierunter fallen die Verwendung von dichotomen oder polytomen Items, die Einschätzung von Testitems oder Testpersonen oder auch die Art des angestrebten Urteilsergebnisses (z.B. Schätzung von Lösungswahrscheinlichkeiten, Personenklassifikationen). Der zweite Bereich betrifft Merkmale des Urteilsprozesses wie die Verwendung von Feedback, Anzahl der Urteilsrunden und die Frage, ob Einzel-

oder Gruppenentscheidungen verlangt werden. Der dritte Bereich bestimmt sich durch die Art des Verfahrens der Cut-Score-Ermittlung, wie z.B. einfache Mittelwertsbildungen, lineare Regressionsmethoden oder die Anwendung von Rasch-Modellen. Kaftandjieva (2004) hat nicht weniger als 34 Methoden nach diesen Unterscheidungsaspekten klassifiziert und beschrieben.

Ein ebenso einfaches wie häufig zitiertes Klassifikationssystem stammt von Jaeger (1989). Dieses System bezieht sich auf einen Teilaspekt der Urteilsaufgabe, d.h. auf die Frage, ob die Beurteiler Einschätzungen hinsichtlich der Testitems oder hinsichtlich der Testpersonen abgeben. Im ersten Fall spricht man von **testzentrierten** Methoden, im zweiten Fall von **personenzentrierten** Methoden.

Bei testzentrierten Methoden schätzen in der Regel 10 bis 20 Beurteiler die Schwierigkeit der einzelnen Testitems ein. Vielfach können diese Einschätzungen erfolgen, noch bevor Testpersonen die Items bearbeitet haben. Demgegenüber zeichnen sich personenzentrierte Methoden durch eine eher holistische Einschätzung der Fähigkeit der Testpersonen aus. Dazu ist es erforderlich, dass die Beurteiler die Testpersonen gut kennen bzw. über relevante Leistungsdaten dieser Personen verfügen. Um die prinzipielle Vorgehensweise in beiden Methodenklassen zu illustrieren, bespreche ich im Folgenden kurz zwei testzentrierte Methoden (die Angoff- und die Bookmarkmethode) sowie zwei personenzentrierte Methoden (die Borderlinegruppen- und die Kontrastgruppenmethode).

Nach der **Angoff-Methode** (Angoff, 1971) werden die zuvor sorgfältig ausgewählten und auf ihre Aufgabe vorbereiteten Beurteiler gebeten, sich eine hypothetische Testperson vorzustellen, deren Fähigkeit exakt an der Schnittstelle zwischen zwei benachbarten Leistungskategorien liegt. Diese Testperson wird üblicherweise „mindestkompetente“ oder „Borderline“-Testperson genannt. Für jedes (dichotome) Testitem haben die Beurteiler dann die Wahrscheinlichkeit (zwischen 0.00 und 1.00) dafür zu schätzen, dass die Borderline-Person eine richtige Antwort gibt. In einer Variante dieser Aufgabe sollen sich die Beurteiler 100 Borderline-Personen vorstellen und schätzen, wie viele dieser Personen das betreffende Item lösen. Die Schätzungen werden pro Beurteiler über alle Items zusammengefasst, um den vom betreffenden Beurteiler erwarteten Score im gesamten Test zu bestimmen. Anschließend werden die erwarteten Testscores über alle Beurteiler gemittelt. Das Ergebnis dieser Mittelung ist der gesuchte Cut-Score.

Bei allen Methoden des Standard-Settings kommt einer möglichst präzisen Definition der jeweils betrachteten Leistungskategorien große Bedeutung zu. Im Falle der Angoff-Methode ist für die beiden benachbarten Kategorien genau festzulegen, welche Leistungsmerkmale die niedrigere und welche Leistungs-

merkmale die höhere Kategorie auszeichnen. Erst vor diesem Hintergrund ist es sinnvoll, das Konzept einer Borderline-Testperson einzuführen. Die zu unterscheidenden Leistungskategorien sind zu benennen ("Performance-Level Labels", PLLs) und jeweils mit Inhalt zu füllen; letzteres kann in Form von "Can-do-Statements" bzw. Kann-Beschreibungen erfolgen ("Performance-Level Descriptors", PLDs). Eines der Ziele des Beurteilertrainings besteht darin, allen Beurteilern soweit wie möglich die gleiche Vorstellung von den Leistungskategorien und den zugehörigen Deskriptoren zu vermitteln.

Die Angoff-Methode ist ein traditionsreiches Verfahren, das mit der Zeit in verschiedener Hinsicht verändert und erweitert wurde. Eine Erweiterung besteht z.B. darin, nicht nur dichotome Items, sondern auch polytome Items betrachten zu können (Hambleton & Plake, 1995). Festzuhalten bleibt allerdings, dass es Beurteilern in der Regel nicht leicht fällt, die geforderten Schätzungen von Lösungswahrscheinlichkeiten mit der gewünschten Genauigkeit abzugeben.

Im Unterschied hierzu ist die **Bookmarkmethode** (Mitzel, Lewis, Patz & Green, 2001) als ein flexibles Verfahren angelegt, das in der Handhabung einfach ist, unterschiedliche Itemformate berücksichtigt und die Festlegung multipler Cut-Scores bei ein und demselben Test zulässt. Nicht zuletzt aufgrund dieser Merkmale hat sich die Bookmarkmethode in den letzten Jahren zu einem der beliebtesten Verfahren des Standard-Settings entwickelt (vgl. Karantonis & Sireci, 2006).

Im Rahmen der Bookmarkmethode werden die Beurteiler gebeten, in einem speziell vorbereiteten Itemheft (dem "Ordered Item Booklet"; kurz OIB) eine oder mehrere Markierungen („Lesezeichen") zu setzen. Das OIB enthält alle Items bzw. Aufgaben des betrachteten Tests in aufsteigender Reihenfolge ihrer Schwierigkeit (jedes Item auf einer separaten Seite). Die dabei zugrunde gelegten Itemschwierigkeiten stammen aus einer vorherigen Analyse des Tests auf der Basis der Item-Response-Theorie (z.B. aus einer Rasch-Analyse). Ein Lesezeichen ist auf derjenigen Seite des OIB anzubringen, bei der die geschätzte Wahrscheinlichkeit dafür, dass eine Borderline-Testperson das betreffende Item richtig beantwortet, erstmals unter .67 fällt, d.h., jedes der vorangegangenen Items beantworten wenigstens zwei von drei Borderline-Personen richtig (vgl. zur Begründung dieses Wertes Mitzel et al., 2001; vgl. auch Huynh, 2006). Für jedes markierte Item wird auf der Grundlage des gewählten Item-Response-Modells das zugehörige Fähigkeitsmaß berechnet. Um schließlich den Cut-Score zu bestimmen, werden die resultierenden Fähigkeitsmaße über die Beurteiler gemittelt.

Bei einer testzentrierten Methode sollen Beurteiler einschätzen, wie **hypothetische** Testpersonen bestimmte Items beantworten. Dagegen verlangt die An-

wendung einer personenzentrierten Methode, dass die Beurteiler den Blick auf **reale** Personen richten, die sie hinsichtlich der vom Test gemessenen Fähigkeit verlässlich einstufen können. Die **Borderlinegruppenmethode** (Livingston & Zieky, 1982) sieht vor, dass die Beurteiler ohne Kenntnis der individuellen Testergebnisse diejenigen Personen benennen, die an der Grenze zwischen zwei benachbarten Kompetenzniveaus liegen (die „Borderlinegruppe"). Alternativ sollen die Beurteiler drei Leistungskategorien bilden: eine Kategorie mit eindeutig kompetenten Personen, eine mit eindeutig nicht kompetenten Personen und eine mit Personen, die zwischen diesen beiden Kategorien einzuordnen wären. In jedem Fall wird die Verteilung der Testscores von Personen aus der Borderlinegruppe ermittelt. Der Median dieser Verteilung bildet den gesuchten Cut-Score.

In der **Kontrastgruppenmethode** (Livingston & Zieky, 1982) sollen die Beurteiler ihnen gut bekannte Personen nach deren wahrgenommener Fähigkeit genau einer von zwei Leistungskategorien zuweisen (z.B. „kompetent" vs. „nicht kompetent"). Im Anschluss an die Durchführung des Tests werden die Score-Verteilungen der beiden Kategorien berechnet. Als Cut-Score ließe sich derjenige Score definieren, an dem sich die Graphen der Verteilungen schneiden. Dieses Vorgehen gewichtet implizit beide Arten von Fehlklassifikationen gleich. Falsch-negative Entscheidungen (eine kompetente Person erhält einen Testscore unterhalb des Cut-Scores) wiegen danach genauso schwer wie falsch-positive Entscheidungen (eine nicht kompetente Person erhält einen Testscore oberhalb des Cut-Scores). Sind die relativen Kosten einer Fehlentscheidung unterschiedlich hoch, dann wäre der endgültige Cut-Score abhängig von den Kosten/Nutzen-Überlegungen nach oben oder unten zu verschieben.

2.4.2. Standard-Setting bei C-Tests

Das Format eines C-Tests stellt besondere Anforderungen an das Standard-Setting. Dies ergibt sich daraus, dass nur der jeweilige Text als Ganzes (im Sinne eines Testlets oder Itembündels; vgl. Wainer, Bradlow & Wang, 2007; Wainer & Kiely, 1987; Wilson & Adams, 1995), nicht aber eine einzelne Lücke als Testitem behandelt werden darf. Hiermit ist das Problem der lokalen Abhängigkeit zwischen Items eines Tests angesprochen (vgl. z.B. Henning, 1989; Wilson, 1988; Yen, 1993).

Die Texte eines C-Tests sind in sich geschlossen und kohärent, und die Lücken innerhalb eines Textes sind inhaltlich, semantisch und syntaktisch eng aufeinander bezogen. Sowohl vorausgegangene Lösungen als auch nachfolgende Teile eines Textes haben daher zwangsläufig erheblichen Einfluss auf die Wahrscheinlichkeit der korrekten Ergänzung einer gegebenen Lücke. Yen (1993) lis-

tete nicht weniger als 10 verschiedene Ursachen lokaler Abhängigkeit zwischen Items eines Tests auf. Wenigstens zwei dieser Ursachen treffen auf die Bearbeitung von C-Test-Texten zu: (a) Abschnittsabhängigkeit ("passage dependence"; dieselbe Information innerhalb einer Textpassage kann zur Lösung verschiedener Items verwendet werden) und (b) Itemketten ("item chaining"; die Lösung eines gegebenen Items erhöht die Chancen, ein anderes Item zu lösen).[5]

Hätten Beurteiler etwa gemäß der Angoff-Methode die Wahrscheinlichkeit dafür einzuschätzen, dass eine Borderline-Testperson eine gegebene Lücke richtig ergänzt, dann müssten sie nicht nur die (wie auch immer konzipierte) Schwierigkeit der Lücke selber, sondern auch alle direkten und indirekten Abhängigkeiten dieser Lücke von allen anderen Lücken desselben Textes berücksichtigen. Dies ist von Beurteilern schwerlich zu leisten. Kaum anders verhielte es sich bei der Bookmarkmethode; hier würden die innerhalb eines Textes bestehenden Abhängigkeitsstrukturen durch die Verwendung des OIB (mit einer Seite pro Lücke) nur besonders augenfällig.

Auf der Ebene von Texten könnte am ehesten die Bookmarkmethode zur Anwendung kommen, allerdings nur unter der Voraussetzung, dass die Anzahl der Texte überschaubar bliebe und nicht zu feine Schwierigkeitsabstufungen resultierten. Ferner müssten je geeignete Deskriptoren vorhanden sein. Diese Voraussetzungen sind aber im Falle des onDaF, der sich auf eine Itembank mit einer großen Menge an skalierten Texten stützt, nicht gegeben. Daher kamen beim onDaF testzentrierte Ansätze des Standard-Settings nicht in Betracht. Stattdessen habe ich ein spezielles Verfahren entwickelt, das der Logik der personenzentrierten Ansätze folgt (siehe Abschnitt 2.3.3).

Ein anderes Problem betrifft die Konstruktion von Fähigkeitsbeschreibungen auf den angestrebten Kompetenzniveaus (d.h. die PLDs). Wie bereits ausgeführt, messen C-Tests allgemeine Sprachkompetenz, nicht jedoch Kompetenz in den einzelnen Sprachfertigkeiten. Die PLDs wären einerseits so allgemein wie nötig, andererseits so klar, anschaulich und nachvollziehbar wie möglich zu formulieren, um die Beurteiler in die Lage zu versetzen, adäquate Vorstellungen von den Niveaus und ihren Grenzziehungen zu entwickeln.

Der onDaF soll, so die Zielsetzung, Einstufungen nicht nur in zwei, sondern in vier Kompetenzniveaus erlauben. Es sind also multiple Cut-Scores festzulegen. Außerdem ist soweit wie möglich sicherzustellen, dass die Niveaubezeich-

[5] Detailanalysen der Schwierigkeit einzelner Lücken, wie sie z.B. Harsch & Schröder (2007) im Rahmen der DESI-Studie vorgenommen haben, sind daher aus psychometrischer Sicht problematisch (vgl. allerdings Harsch & Hartig, im vorliegenden Band).

nungen (PLLs) und die Niveaubeschreibungen (PLDs) einem einheitlichen, allgemein akzeptierten und leicht verständlichen System folgen.[6]

Im Falle des onDaF fiel die Wahl auf das System des GER. Da die GER-Stufenbeschreibungen je verschiedene Ausprägungen von Fähigkeiten in den vier Sprachfertigkeiten wiedergeben, ist unmittelbar einsichtig, dass die Beziehung der onDaF-Stufen zum GER nur indirekt sein kann.

Die empirische Basis für die Annahme, dass sich eine indirekte Beziehung zum GER herstellen lässt, bilden Untersuchungen zur Konstruktvalidität von C-Tests (Eckes & Grotjahn, 2006a). In Abschnitt 4.1 gehe ich genauer auf die Ergebnisse dieser Untersuchungen ein. Hier sei nur soviel vorweggenommen: C-Tests besitzen statistisch hochsignifikante Zusammenhänge mit separaten Tests der rezeptiven Fertigkeiten (Lesen, Hören) und der produktiven Fertigkeiten (Schreiben, Sprechen). Diese Zusammenhänge gehen auf den Faktor der allgemeinen Sprachkompetenz zurück. Mit anderen Worten, je höher die allgemeine Sprachkompetenz einer Person ausgeprägt ist, desto größer sind ihre Chancen, bei einem fertigkeitsspezifischen Test ein gutes Resultat zu erzielen.

2.4.3. Standard-Setting beim onDaF: Die Prototypgruppenmethode

Das Hauptziel der beim onDaF verwendeten Methode des Standard-Settings lässt sich wie folgt umreißen: Festlegung multipler Cut-Scores, die eine Einstufung der allgemeinen Sprachkompetenz von Testpersonen analog zur globalen Skala des GER erlauben. Der Wertebereich der Skala wurde dabei aus mehreren Gründen eingeschränkt: (a) die Zielgruppe bilden ausländische Studieninteressierte, die wenigstens über Grundkenntnisse des Deutschen verfügen, (b) Personen mit überragenden Deutschkenntnissen sind weniger an einem Einstufungstest interessiert, als an einem Test, der ihre Sprachkompetenz nach Fertigkeiten getrennt misst (wie dies der TestDaF leistet), (c) es ist relativ aufwändig, sehr leichte bzw. sehr schwere C-Test-Texte in großer Zahl nach demselben Prinzip zu konstruieren, und (d) die Dauer des Einstufungstests sollte deutlich weniger als eine Stunde betragen.

Der onDaF zielt daher darauf ab, Einstufungen analog zu den Stufen A2, B1, B2 und C1 der globalen Skala des GER zu erlauben; unterhalb von A2 und oberhalb von C1 differenziert der Test nicht. Um die Stufe A2 eindeutig zu bestim-

---

[6] Die Verbindung von Testergebnissen in einem C-Test mit Niveaus der Sprachkompetenz ist kein einfaches Unterfangen. Bisherige Lösungsansätze wie solche, die sich auf eine Klassifikation von Lehrwerktexten (Baur & Spettmann, 2006), auf eine Segmentierung von Testscoreverteilungen (Zydatiß, 2005; Vockrodt-Scholz & Zydatiß, im vorliegenden Band) oder auf eine lineare Transformation äquidistanter Intervallgrenzen (Reichert, Keller & Martin, im vorliegenden Band) stützen, verdeutlichen die methodischen Schwierigkeiten.

men, ist es allerdings notwendig, einen „unteren“ Cut-Score für A2, d.h. einen Cut-Score zwischen A2 und allen niedrigeren Kompetenzbereichen, zu ermitteln. Die onDaF-Stufe C1 ist dagegen nach oben offen; sie steht für alle Kompetenzniveaus oberhalb von B2. Diese Stufe lässt sich als eine Abkürzung für „C1 oder höher“ verstehen; ein oberer Cut-Score ist nicht erforderlich.

Der beim onDaF eingeschlagene Weg zur Festlegung der Cut-Scores basiert, wie bereits ausgeführt, auf einer personenzentrierten Methodik. Diese Methodik vermeidet es zwar, den Beurteilern die nicht einfach zu realisierende Vorstellung von hypothetischen Borderline-Testpersonen und ihrem wahrscheinlichen Antwortverhalten abzuverlangen, ist aber ihrerseits an eine Reihe von Voraussetzungen geknüpft (Zieky & Perie, 2006). So dürfen die Beurteilungen nur von hierfür qualifizierten Personen vorgenommen werden, d.h., es kommen nur solche Personen als Beurteiler in Betracht, die sowohl in der Einschätzung der fraglichen Eigenschaft kompetent sind als auch die zu beurteilenden Personen hinsichtlich dieser Eigenschaft gut kennen. Weiter müssen die Beurteilungen auf genau jene Eigenschaft zielen, die der Test messen soll, in großer zeitlicher Nähe zum Test erfolgen und die unverfälschten Meinungen der Beurteiler über die Ausprägungen der Eigenschaft wiedergeben, d.h., die Beurteiler dürfen nicht motiviert sein, besonders streng oder milde zu urteilen.

Sind diese Voraussetzungen als gegeben zu betrachten, so könnte die Borderlinegruppen- oder die Kontrastgruppenmethode Anwendung finden. Allerdings stellen auch diese Methoden Anforderungen an die Beurteiler, die nicht unerheblich sind. Häufig ist nicht ausreichend klar, welche Personen zur Borderlinegruppe zählen bzw. welche Personen den beiden Kontrastgruppen angehören und wie diese Gruppen adäquat zu definieren sind. Die Schwierigkeiten werden noch größer, wenn multiple Cut-Scores bestimmt werden sollen.

Ein anderer Ansatz geht von der Idee aus, das Problem des Standard-Settings als ein **Klassifikationsproblem** zu konzipieren (vgl. Sireci, 2001): Personen sind nach bestimmten Merkmalen in eine Reihe von Gruppen oder Klassen aufzuteilen, die intern möglichst homogen und extern (voneinander) möglichst separiert sind. Die Grenzen zwischen benachbarten Klassen bilden die Cut-Scores. Auf der Basis dieser Überlegungen entwickelte Sireci (2001; vgl. auch Sireci, Robin & Patelis, 1999) einen clusteranalytischen Ansatz des Standard-Settings.

Das Klassifikationskonzept des Standard-Settings erlaubt es, eine Brücke zur psychologischen Begriffs- und Kategorisierungsforschung zu schlagen (vgl. für Übersichtsdarstellungen z.B. Eckes, 1991, 1996; Murphy, 2004; vgl. auch Haswell, 1998). Unter den oben genannten Bedingungen der Anwendung personenzentrierter Methoden lässt sich postulieren, dass Beurteiler mit zunehmender Berufserfahrung (explizit oder implizit) eine recht klare Vorstellung davon entwi-

ckeln, welche Personen typische Vertreter ihrer jeweiligen Kategorie sind, also z.B. typische Vertreter der Kategorie der leistungsstarken oder leistungsschwachen Deutschlerner.

Der typische Vertreter oder das beste Beispiel einer Kategorie wird in der Begriffsforschung **Prototyp** genannt. Allgemein gesprochen ist ein (kognitiver) Prototyp eine mentale Repräsentation der zentralen Tendenz einer Kategorie von Entitäten (z.B. Objekte, Personen, Situationen; vgl. Barsalou, 1992; Eckes, 1996). Prototypen bieten eine Reihe von Vorteilen für die menschliche Informationsverarbeitung. Sie sind u.a. rascher und genauer zu erinnern und zu erkennen und liefern konsistentere Einschätzungen ihrer Merkmale als andere Vertreter einer Kategorie, und zwar insbesondere als solche Vertreter, die an der Grenze zwischen zwei Kategorien liegen (Eckes, 1996; Murphy, 2004).

Die eigens für den onDaF entwickelte Standard-Setting-Methode stützt sich wesentlich auf das Prototypkonzept. Diese Methode, sie sei im Folgenden **Prototypgruppenmethode** genannt, sieht vor, Sprachlehrer bzw. Leiter von Sprachkursen diejenigen Deutschlerner benennen zu lassen, die sie als typische Vertreter einer definierten Leistungsstufe betrachten. Dabei erfolgt die Definition der angestrebten Kompetenzstufen A2, B1, B2 und C1 anhand von Kann-Beschreibungen, die der globalen GER-Skala entnommen sind (Europarat, 2001, S. 35). Die Sprachlehrer werden also gebeten, den typischen Lerner auf der Stufe A2 (den „A2-Prototyp“), den typischen Lerner auf der Stufe B1 (den „B1-Prototyp“) usw. anzugeben. Voraussetzung hierfür ist, dass sie über genügend Information verfügen, um den Leistungsstand ihrer jeweiligen Lerner oder Kursteilnehmer zuverlässig nach den GER-Kriterien einschätzen zu können.

Die Lernerprototypen werden auf der Basis einer gemeinsamen Kalibrierung aller Testpersonen und Texte auf der Logitskala lokalisiert. Das heißt, es werden die Fähigkeitsschätzungen (Logits) der als Prototypen ausgewählten Teilnehmer identifiziert und anhand der Einstufungen entlang der GER-Skala (A2 bis C1) klassifiziert. Um auf der Basis dieser Logit-Klassifikationen Cut-Scores zu ermitteln, kommen zwei verschiedene statistische Verfahren in Betracht.

Das erste Verfahren definiert Cut-Scores anhand der Mediane der Verteilungen von Logitwerten, die im Überlappungsintervall je zweier benachbarter Kompetenzstufen (Leistungskategorien) liegen. Das zweite Verfahren stützt sich auf das Modell der binären logistischen Regression (Livingston & Zieky, 1989). Dieses Regressionsmodell erlaubt (vereinfacht gesprochen) die Schätzung desjenigen Testscores, der am besten zwischen zwei benachbarten Kategorien (hier z.B. zwischen A2- und B1-Lernern) trennt; der Score mit der höchsten „Trennschärfe“ ist der gesuchte Cut-Score. Eine ausführlichere Darstellung der Prototypgruppenmethode und ihrer Ergebnisse findet sich in Abschnitt 3.2.

3. Konstruktion

Das übergeordnete Ziel der Testentwicklung bestand, wie schon mehrfach erwähnt, im sukzessiven Aufbau einer kalibrierten Itembank. Diese Itembank sollte eine stabile, internetgestützte Durchführung des onDaF ermöglichen.

Die fünf wesentlichen Entwicklungsschritte waren: (a) Erstellung von Lückentexten nach dem C-Test-Prinzip, (b) Erprobung der Texte in Sets zu je 10 Texten, inklusive zweier Ankertexte, (c) separate Rasch-Analyse jedes einzelnen Sets und Ausschluss psychometrisch ungeeigneter Texte, (d) Rasch-Analyse der Gesamtmenge verbliebener Texte zur Kalibrierung ihrer Schwierigkeiten auf einer gemeinsamen Skala und zur Ermittlung der Lernerprototypen für das Standard-Setting, (e) Eingabe der Texte mit ihren Attributen in die Itembank.

Der vorliegende Abschnitt informiert über die umfangreichen Datenerhebungen, die Ergebnisse der Rasch-Analysen und der verschiedenen Analysen zum Standard-Setting. Im Mittelpunkt stehen die Erprobungen der Texte an Gruppen von Testpersonen aus der Population von potenziellen onDaF-Teilnehmern, die Anordnung der Texte auf einer gemeinsamen Schwierigkeitsskala und die Festlegung der Cut-Scores nach der Prototypgruppenmethode.

3.1. Erprobung von C-Test-Texten

3.1.1. Teilnehmer

Die hier berichteten Erprobungen sind Teil eines fortlaufenden Prozesses der Erstellung, Evaluation und Kalibrierung von C-Test-Texten. Im Folgenden betrachte ich die ersten 18 Erprobungen, durchgeführt im Zeitraum zwischen März 2005 und Februar 2006. Spätere Erprobungen bleiben unberücksichtigt.

Es nahmen insgesamt 3.651 Personen teil, darunter 2.270 Frauen (62.2%) und 1.364 Männer (37.4%). Keine Angaben zum Geschlecht machten 17 Personen. Das Alter von rund 80% der Teilnehmer lag zwischen 18 und 27 Jahren (M = 23.43, SD = 5.91).

Die Teilnehmer (auch Probanden, kurz Pbn) stammten aus 116 Ländern. Am häufigsten vertreten waren die folgenden 10 Herkunftsländer (in Klammern die Teilnehmerzahl): Russische Föderation (453), Volksrepublik China (240), Ukraine (175), Marokko (154), Bulgarien (153), Finnland (143), Türkei (139), Frankreich (138), Tschechien (136), Uganda (133). Die Pbn verteilten sich auf 92 Studienfächer bzw. Studienfelder; der weit überwiegende Teil studierte Germanistik (1.133), gefolgt von Wirtschaftswissenschaften (331), Informatik (114) und Maschinenbau/Fahrzeugtechnik (65); zur Schule gingen noch 165 Pbn. Zum Zeitpunkt der Datenerhebung besuchten die Pbn Sprachkurse an Lektoraten des Deutschen Akademischen Austauschdienstes (DAAD) oder an Testzentren des

TestDaF-Instituts. Es beteiligten sich Lektorate bzw. Testzentren aus 38 Ländern, verteilt über fünf Kontinente.

Alle Pbn nahmen freiwillig und unentgeltlich an den Erprobungen teil. Nach Abschluss der Datenanalysen erhielt jeder Pb eine Rückmeldung über die erreichte Punktzahl und über seinen Prozentrang (bezogen auf die jeweilige Erprobungsgruppe).

3.1.2. Testmaterial

Die zu erprobenden Lückentexte waren nach dem klassischen Konstruktionsprinzip erzeugt worden (vgl. z.B. Grotjahn, 2002). Jeder Text enthielt 20 Lücken. Die einzelnen Lücken erschienen als durchgehende Striche konstanter Länge. Jeder Text war mit einer knappen Überschrift versehen. Beispiele für die verwendete Art von Texten finden sich auf der onDaF-Homepage unter www.ondaf.de, Link „Beispieltest“ (vgl. auch Eckes, 2006b).

Für jede Erprobung wurde ein Set von jeweils 10 Texten gebildet. Innerhalb eines Sets waren die Texte nach aufsteigender Schwierigkeit angeordnet; jeder Text behandelte ein anderes Thema. Die Schwierigkeit der (noch unskalierten) Texte wurde anhand von Vorerprobungen und durch Expertenurteile bestimmt.

Dem Konzept eines Ankertestplans folgend wurden an der dritten und achten Position zwei sorgfältig ausgewählte Texte platziert, die als Ankertexte fungierten. Das heißt, diese beiden Texte waren allen Erprobungssets gemeinsam und sollten so eine Verbindung zwischen den verschiedenen Erprobungssets herstellen. Die Gesamtzahl der psychometrisch untersuchten Texte belief sich demzufolge auf 146.

In den Erprobungen wurden ausschließlich Papierversionen von C-Test-Texten verwendet. Jeder Text erschien auf einer separaten Seite eines Testhefts. Auf der ersten Seite des Testhefts wurden die Pbn gebeten, einige Angaben zur Person zu machen (Alter, Geschlecht, Herkunftsland, Studienfach). Diese Seite enthielt auch eine kurze Instruktion zur Bearbeitung des Tests.

Die Testleiter (DAAD-Lektoren oder TestDaF-Prüfungsbeauftragte) bekamen zusätzlich zu den Testheften einen Fragebogen, der hauptsächlich den Zweck verfolgte, die für die Prototypgruppenmethode relevanten Informationen zu gewinnen. Zunächst sollten die Testleiter einschätzen, welche Stufen der globalen Skala des GER in ihrer Teilnehmergruppe vertreten waren. Die Stufenskala reichte von A1 (ganz elementare Sprachverwendung) bis C2 (hoch kompetente Sprachverwendung). Mehrfachnennungen waren möglich. Zur Erläuterung der Fertigkeiten, die auf den einzelnen Stufen vorhanden sein sollten, war ein Abdruck von Tabelle 1 der GER-Publikation (Europarat, 2001, S. 35) beigefügt.

Anschließend sollten die Testleiter einzelne Teilnehmer, die sie hinsichtlich ihrer Deutschkenntnisse sehr gut beurteilen konnten, in den Blick nehmen und jene Teilnehmer benennen, die nach ihrem Eindruck besonders typisch für Sprachkompetenz auf den Stufen A2, B1, B2 oder C1 des GER waren. Das heißt, sie sollten angeben, welche Teilnehmer aus ihrer Sicht als typische A2-Lerner, welche als typische B1-Lerner usw. einzuschätzen waren. Pro Stufe sollten sie die Namen von maximal drei Teilnehmern in eine Liste eintragen.

3.1.3. Durchführung und Auswertung

Jeder Pb erhielt ein Testheft mit der Instruktion auf der ersten Seite und je einem Lückentext auf den folgenden 10 Seiten. Pro Text standen genau fünf Minuten Bearbeitungszeit zur Verfügung. Wie bei C-Tests zumeist üblich, erhielten die Pbn lediglich die Anweisung, die Lücken in sinnvoller Weise zu ergänzen. Die Texte waren ausschließlich in der vorgegebenen Reihenfolge zu bearbeiten, ein Zurück- oder Vorblättern war nicht erlaubt.

Die Teilnehmerantworten, d.h. die Ergänzungen der Textlücken, wurden wie folgt ausgewertet: Als „korrekt“ galten allein orthografisch richtige Originale oder orthografisch richtige Varianten (Scoringmethode A in Eckes & Grotjahn, 2006b). Für jede korrekte Ergänzung wurde genau ein Punkt vergeben. Jede inkorrekte oder fehlende Ergänzung wurde mit 0 Punkten bewertet. Maximal waren 200 Punkte zu erreichen. Andere Auswertungsverfahren, wie z.B. jene, die auch orthografisch falsche Originale als korrekt gelten lassen, kamen nicht zum Zuge, da sie sich in einer früheren Untersuchung (Eckes & Grotjahn, 2006b) als fehleranfälliger erwiesen hatten (vgl. zu möglichen Auswertungsverfahren auch Cronjäger, Klapheck, Krätzschmar & Walter, im vorliegenden Band).

3.1.4. Deskriptive Statistiken

Tabellen 1a und 1b geben eine Übersicht über die deskriptiven Statistiken zu den 18 Erprobungen (abgekürzt als E01 bis E18).

Die Stichprobenumfänge können mit Werten zwischen 168 Pbn (E01, E07) und 276 Pbn (E03) als ausreichend hoch betrachtet werden; durchschnittlich nahmen 203 Pbn teil. Auffällig sind die Unterschiede in den Mittelwerten der Score-Verteilungen. Während 10 Erprobungen Mittelwerte um 100 Punkte aufweisen, liegen die Mittelwerte bei 3 Erprobungen (E05, E06 und E08) über 120 Punkte und bei 5 Erprobungen (E07, E11, E14 bis E16) zum Teil klar unter 80 Punkte. Die größte Mittelwertsdifferenz beträgt rund 61 Punkte (zwischen E06 und E16). Auch die Streuungen der Score-Verteilungen schwanken erheblich.

Die mit einigem Abstand größten Streuungen liegen für E02 und E11 vor (Werte über 46), den kleinsten Streuungswert besitzt E07 (mit 27.4).

Tabelle 1a: Deskriptive Statistiken und summarische Rasch-Statistiken für Erprobungen E01 bis E09

	E01	**E02**	**E03**	**E04**	**E05**	**E06**	**E07**	**E08**	**E09**
N	168	187	276	189	196	204	168	188	186
	Deskriptive Statistiken								
***M* (Score)**	114.4	109.0	104.4	108.5	121.8	124.3	77.6	124.2	108.0
***SD* (Score)**	34.5	46.6	45.1	44.2	40.0	35.4	27.4	32.8	37.7
Max.	198	192	196	193	197	192	150	193	190
Min.	38	9	8	19	24	15	6	36	22
	Summarische Rasch-Statistiken								
***M* (Logit)**	.29	.30	.10	.31	.64	.54	–.55	.61	.22
***SD* (Logit)**	.88	1.16	1.25	1.33	1.11	.89	.78	.87	.93
***SE* (Logit)**	.16	.17	.17	.18	.17	.16	.17	.17	.16
Klassensep.	6.45	8.49	8.69	9.20	7.73	6.61	5.69	6.59	7.31
Reliabilität	.95	.97	.98	.98	.97	.96	.94	.96	.96

Anmerkung: Deskriptive Statistiken beziehen sich auf die jeweils erreichten Testwerte (Summenscores). Rasch-Statistiken beziehen sich auf die Schätzungen der Personenparameter (Personenfähigkeit) in Einheiten der Logitskala. *M* = arithmetischer Mittelwert. *SD* = Standardabweichung. *SE* = Standardfehler der Parameterschätzung. Der Index der Klassenseparation gibt die Anzahl statistisch reliabel unterscheidbarer Klassen von Pbn an. Die Reliabilität ist die Pbn-Separationsreliabilität (entspricht Cronbachs Alpha).

Da es sich bei den Pbn-Gruppen um nichtäquivalente Gruppen handelt, ist die Schwierigkeit eines Erprobungssets mit der Fähigkeit der Pbn, die dieses Set zu bearbeiten hatten, konfundiert. Mit anderen Worten, es lässt sich ohne Berücksichtigung zusätzlicher Informationen nicht erkennen, ob beispielsweise im Falle von E06 (Score-Mittelwert 124.3) die Texte besonders leicht oder die Pbn besonders fähig waren (oder beides). Umgekehrt könnten im Falle von E16 (Score-Mittelwert 63.5) eher durchschnittlich fähige Pbn besonders schwere Texte vor sich gehabt haben. Eine Trennung zwischen Text-Schwierigkeit und Pbn-Fähigkeit ist jedoch anhand der Ankertexte im Rahmen von Rasch-Analysen möglich.

Tabelle 1b: Deskriptive Statistiken und summarische Rasch-Statistiken für Erprobungen E10 bis E18

	E10	**E11**	**E12**	**E13**	**E14**	**E15**	**E16**	**E17**	**E18**
N	188	222	177	212	234	230	208	206	212
	Deskriptive Statistiken								
***M* (Score)**	96.6	72.3	104.5	104.9	66.3	71.9	63.5	104.3	102.4
***SD* (Score)**	32.1	47.0	38.8	35.1	32.5	31.7	31.1	33.8	38.3
Max.	189	198	173	182	162	158	176	175	190
Min.	18	4	16	30	13	10	10	31	13
	Summarische Rasch-Statistiken								
***M* (Logit)**	–.08	–.77	.17	.31	–.89	–.83	–.90	.17	.10
***SD* (Logit)**	.83	1.40	1.03	.89	.98	.90	.94	.90	1.01
***SE* (Logit)**	.16	.19	.17	.16	.18	.17	.18	.17	.17
Klassensep.	6.41	8.87	7.77	6.95	6.60	6.35	6.40	6.87	7.59
Reliabilität	.95	.98	.97	.96	.96	.95	.95	.96	.97

Anmerkung: Deskriptive Statistiken beziehen sich auf die jeweils erreichten Testwerte (Summenscores). Rasch-Statistiken beziehen sich auf die Schätzungen der Personenparameter (Personenfähigkeit) in Einheiten der Logitskala. *M* = arithmetischer Mittelwert. *SD* = Standardabweichung. *SE* = Standardfehler der Parameterschätzung. Der Index der Klassenseparation gibt die Anzahl statistisch reliabel unterscheidbarer Klassen von Pbn an. Die Reliabilität ist die Pbn-Separationsreliabilität (entspricht Cronbachs Alpha).

Im folgenden Abschnitt berichte ich zunächst über Rasch-Analysen, die getrennt nach Erprobungen durchgeführt wurden. Anschließend gehe ich auf die für die Entwicklung der kalibrierten Itembank zentrale Rasch-Analyse ein. Diese Analyse berücksichtigte alle Erprobungen gleichzeitig und erlaubte so die Konstruktion einer gemeinsamen Skala der Schwierigkeiten sämtlicher Texte.

3.1.5. Separate Rasch-Analysen

Wie bereits angesprochen, sind die Lücken innerhalb eines C-Test-Textes voneinander abhängig, sodass jeder Lückentext als Ganzes, d.h. als Testlet oder Itembündel, aufzufassen und zu analysieren ist (siehe Abschnitt 2.3.2). C-Test-Texte bilden so gesehen polytome Items mit $m + 1$ Itemwerten oder, anders ausgedrückt, Ratingskalen mit $m + 1$ Kategorien, wobei die sukzessiven Kategorien k ganzzahlig kodiert sind (d.h. $k = 0, 1, \ldots, m$); hierbei ist m die Anzahl der Lücken eines Textes. Im vorliegenden Fall hat jeder Text genau 20 Lücken. Die Texte können 21 verschiedene Werte (d.h. Werte im Intervall zwischen 0 und 20) annehmen.

Innerhalb der Item-Response-Theorie (IRT) gibt es mehrere Testmodelle, die eine Skalierung von gestuften polytomen Items erlauben (vgl. z.B. Embretson & Reise, 2000; Müller, 1999; Ostini & Nering, 2006; Rost, 2004). In der vorliegenden Arbeit kam das diskrete Ratingskalenmodell (RSM; Andrich, 1978, 1982) zum Einsatz. Dieses Modell hatte sich in vorhergehenden Untersuchungen zur Rasch-Skalierung von C-Tests bewährt (Eckes, 2006b, 2007). Zudem war ein Computerprogramm zur Anwendung des RSM verfügbar, das sich für eine gemeinsame Kalibrierung von C-Test-Texten aus verschiedenen Testformen mit je anderen Pbn-Gruppen eignete. Es handelte sich um das Programm WINSTEPS (Version 3.64; Linacre, 2007).

Das RSM setzt voraus, dass alle Items dasselbe Antwortformat haben. Diese Voraussetzung ist bei einem C-Test, der nach dem klassischen Tilgungsprinzip erstellt wurde, und für den gilt, das alle Texte dieselbe Anzahl von Lücken aufweisen, als gegeben zu betrachten. Für alle derartigen Items schätzt das RSM eine gemeinsame Menge von Schwellenparametern. Das heißt, den Antwortkategorien werden Schwellenparameter zugewiesen, die bis auf Unterschiede in den Itemschwierigkeiten, also bis auf unterschiedliche Lokationen der Items auf dem latenten Kontinuum, für alle Items identisch sind. Anders ausgedrückt, Unterschiede zwischen den Items (Texten) bestehen allein hinsichtlich ihrer Schwierigkeit, nicht aber hinsichtlich der Schwierigkeit der Übergänge zwischen den einzelnen Antwortkategorien (Lücken) innerhalb der Texte.[7]

In der unteren Hälfte von Tabellen 1a und 1b finden sich die Ergebnisse der Rasch-Analysen in Form allgemeiner, summarischer Statistiken. Dabei ist zu beachten, dass die Statistiken auf separaten Rasch-Analysen der Daten aus den Erprobungen basieren, d.h., es liegt diesen Werten keine gemeinsame Skalierung der Texte zugrunde. Die mittlere Schwierigkeitsschätzung der Texte innerhalb einer Erprobung wurde wie üblich zur Festlegung des Ursprungs der Logitskala zentriert, also gleich Null gesetzt.

Die mittleren Logitwerte der Personenparameter spiegeln die auf der Ebene der Testscores beobachtete Unterschiedlichkeit der Erprobungsgruppen wider. Gleiches gilt für die Standardabweichungen der Parameterschätzungen. Die Werte des Standardfehlers als Maß der Genauigkeit der Parameterschätzungen liegen im Vergleich der Erprobungen auf einem konsistent niedrigen Niveau.

Der Index der Klassenseparation (vorletzte Zeile von Tabelle 1) gibt die Anzahl der statistisch reliabel unterscheidbaren Klassen von Pbn an (vgl. Eckes, 2006b, 2007; Wright & Masters, 1982, 2002). Ein Wert von 6.45 für E01 besagt

---

[7] Unterschiedlich schwierige Übergänge zwischen Antwortkategorien lassen sich prinzipiell nach dem Partial-Credit-Modell (Masters, 1982) analysieren. Dies erschien aber im vorliegenden Kontext nicht angezeigt (vgl. für eine ausführliche Diskussion Eckes, 2006b).

z.B., dass sich die Stichprobe von Pbn, die an der Erprobung E01 teilgenommen haben, anhand der Schätzungen ihrer Fähigkeitsparameter in etwa sechseinhalb statistisch signifikant voneinander getrennte Klassen einteilen lässt. Die Werte der Klassenseparation schwanken zwischen 5.69 (E07) und 9.20 (E04). Das heißt, in keiner einzigen Erprobung sind weniger Klassen von Pbn zuverlässig unterscheidbar, als der onDaF Kompetenzstufen erfassen soll. In den meisten Erprobungen ergeben sich Separationswerte, die um wenigstens zwei Klassen über dem notwendigen Minimum (4 Klassen bzw. Stufen) liegen.

Der Index der Separationsreliabilität (letzte Zeile von Tabellen 1a und 1b) erfasst die Genauigkeit, mit der die Fähigkeitsmaße voneinander unterschieden werden können. Dieser Index entspricht einer Reliabilitätsschätzung nach Cronbachs Alpha. Die Reliabilität liegt mit Werten zwischen .94 und .98 auf einem sehr hohen Niveau.

Eine andere wichtige Frage bei der Anwendung eines Rasch-Modells ist die nach der Modellgültigkeit. Die Gültigkeit eines Modells wie des RSM kann u.a. anhand der Reproduzierbarkeit der Daten überprüft werden (vgl. z.B. Bond & Fox, 2007; Eckes, 2006b; Müller, 1999). WINSTEPS liefert zu jedem Item Informationen darüber, wie gut die Daten den Erwartungen des Messmodells entsprechen bzw. wie gut das Modell die Daten reproduzieren kann. Diese Informationen werden in Form so genannter Mean-Square-Fit-Statistiken zusammengefasst („Infit“, „Outfit“; Wright & Masters, 1982; vgl. auch Eckes, 2005).

Infit- und Outfit-Statistiken haben einen Erwartungswert von 1; sie können Werte im Bereich zwischen 0 und +∞ annehmen. Werte deutlich größer 1 verweisen darauf, dass die Daten anhand des Modells nicht gut vorhersagbar sind, oder anders ausgedrückt, nur schlecht mit den Modellannahmen übereinstimmen (“Misfit”, “Underfit”) bzw. mehr Variation aufweisen, als es den Erwartungen des Modells entspricht. Dies wäre z.B. der Fall, wenn ein eher leistungsschwacher Pb bei relativ schweren Texten hohe Punktzahlen erreicht. Umgekehrt indizieren Werte deutlich kleiner 1, dass ein relativ hohes Maß an Vorhersagbarkeit oder Redundanz vorliegt, die Daten „zu gut“ auf das Modell passen (“Overfit”) bzw. weniger Variation zeigen als vorhergesagt. Allgemein ist Overfit weniger problematisch als Misfit (Myford & Wolfe, 2003). Der Unterschied zwischen beiden Statistiken liegt darin, dass bei der Infit-Statistik extreme Abweichungen von den Modellerwartungen weniger ins Gewicht fallen als bei der Outfit-Statistik.

Linacre (2002) hat grobe Richtwerte für die Interpretation von Mean-Square-Statistiken vorgeschlagen. Danach sind Fitwerte im Intervall zwischen 0.5 und

1.5 messmethodisch akzeptabel.[8] Liegen die Werte der Fit-Statistiken deutlich außerhalb dieser Intervallgrenzen, wäre von einer mangelnden „lokalen“ Modellanpassung zu sprechen. „Lokal“ bedeutet, dass anders als bei einem globalen Test auf Modellgültigkeit nicht das Modell als Ganzes, sondern nur bestimmte Teile des Modells, eben jene, die mit den erhöhten Fitwerten verbunden sind, nicht mit den Daten in Einklang stehen. Solche Modellverletzungen wären stets genauer zu untersuchen. Sie könnten z.B. darauf zurückzuführen sein, dass der Test nicht eindimensional ist.

Da das RSM ein eindimensionales Modell ist, d.h. ein Modell, das die Daten durch eine einzige latente Dimension zu erklären versucht, ist in der Fit-Analyse zugleich eine Prüfung der Eindimensionalität eines gegebenen Tests zu sehen. Tabelle 2 fasst für unterschiedlich eng definierte Intervalle die Ergebnisse der Fit-Analysen in den einzelnen Erprobungen zusammen.

In keinem einzigen Fall überschreiten die Werte der beiden Fit-Statistiken die Grenzen des 0.50/1.50-Intervalls. Nahezu alle Werte liegen noch innerhalb des enger definierten 0.70/1.30-Intervalls. Etwa die Hälfte aller Werte fällt in das sehr eng definierte 0.90/1.10-Intervall; in diesem Intervall ist außerdem der Anteil der “Misfits” um ca. die Hälfte kleiner als der Anteil der “Overfits”. Insgesamt kann daher von einer guten Modellanpassung gesprochen werden. Die Ergebnisse der Fit-Analysen stützen damit die Modellannahme der Eindimensionalität der Texte innerhalb der Erprobungssets.

Tabelle 2: Häufigkeit von Infit- und Outfit-Werten für Texte aus Erprobungssets bei unterschiedlich weiten Fit-Intervallen

	Infit		**Outfit**	
Intervall	**Häufigkeit**	**%**	**Häufigkeit**	**%**
Fit < 0.50	0	0.0	0	0.0
0.50 ≤ Fit ≤ 1.50	180	100	180	100
Fit > 1.50	0	0.0	0	0.0
Fit < 0.70	4	2.2	2	1.1
0.70 ≤ Fit ≤ 1.30	174	96.7	173	96.1
Fit > 1.30	2	1.1	5	2.8
Fit < 0.90	67	37.2	63	35.0
0.90 ≤ Fit ≤ 1.10	87	48.3	94	52.2
Fit > 1.10	26	14.4	23	12.8

Anmerkung: Die Häufigkeitsangaben beziehen sich auf 18 Erprobungssets mit je 10 Texten.

[8] Je nach Fragestellung oder Verwendungszusammenhang der Untersuchungsergebnisse können die Intervalle auch breiter oder enger definiert werden (vgl. Bond & Fox, 2007, S. 238–243).

Ein weiterer „Prüfstein“, dem sich C-Test-Texte stellen müssen, bevor sie für eine Aufnahme in die Itembank in Frage kommen, ist die Analyse differenzieller Itemfunktionen (DIF-Analyse). Wie schon bei der Besprechung allgemeiner Testgütekriterien (siehe Abschnitt 2.1) erwähnt, lässt sich in DIF-Analysen feststellen, ob einzelne Items für verschiedene Gruppen von Testpersonen unterschiedlich schwierig sind, und zwar bei gleicher durchschnittlicher Fähigkeit der Gruppen.

Da DIF-Analysen relativ hohe Anforderungen an die Anzahl von Personen pro Gruppe stellen, wird routinemäßig nur das Geschlecht der Teilnehmer an den Erprobungen betrachtet. Die Gender-DIF-Analysen erfolgten auf der Basis der RSM-Parameterschätzungen nach dem bei Linacre (2007) beschriebenen Verfahren. Unter Berücksichtigung der Bonferroni-Adjustierung zur Korrektur des Alpha-Fehlers führten diese Verfahren zum Ausschluss von 10 Texten. Vier Texte waren für Frauen schwieriger als für Männer, bei sechs Texten verhielt es sich umgekehrt. In keinem Fall war ein inhaltlicher Aspekt für die differenzielle Funktionsweise der kritischen Texte auszumachen.

3.1.6. Simultane Rasch-Analyse

Um alle Texte aus den 18 Erprobungen (ausgenommen die Texte mit auffälligen Gender-DIF-Werten) auf einer gemeinsamen Schwierigkeitsskala zu kalibrieren, fügte ich die einzelnen Datensätze zu einem einzigen Datensatz zusammen. Der Gesamtdatensatz bestand aus den Scores, die 3.651 Teilnehmer bei 135 Texten erreicht hatten (1 Text war aufgrund eines Fehlers bei der Zusammenstellung der Erprobungssets zweifach verwendet worden). Als Bindeglied zwischen den separaten Datensätzen dienten die beiden Ankertexte. Abbildung 3 zeigt die von WINSTEPS (Linacre, 2007) erzeugte gemeinsame Verteilung der Schätzungen von Pbn- und Textparametern.

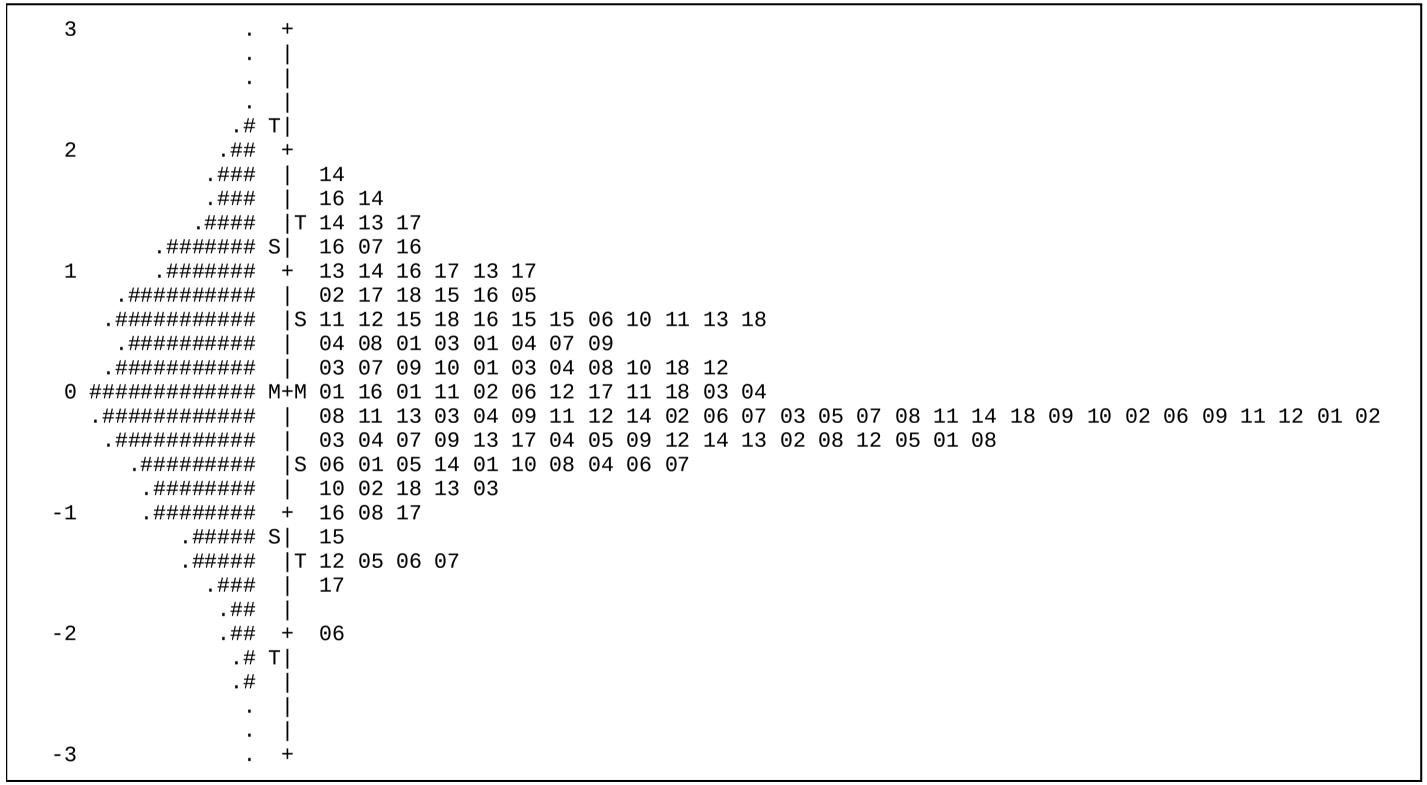

Abbildung 3: Gemeinsame Darstellung der Fähigkeitsparameter von 3651 Pbn (linker Teil) und der Schwierigkeitsparameter von 135 Texten (rechter Teil) ausgedrückt in Einheiten der Logitskala (ganz links). Ein Rautenzeichen („#“) steht für 22 Pbn, ein Punkt für 1 bis 21 Pbn. Texte werden durch die Nummer des Erprobungssets, dem sie zugeordnet sind, bezeichnet.

Ganz links in dieser Abbildung ist die Logitskala wiedergegeben. Höhere Logitwerte zeigen eine größere Fähigkeit der Pbn bzw. eine größere Schwierigkeit der Texte an. Aus Platzgründen habe ich die Logitskala, die in dieser Analyse von –5 bis 5 reichte, an den Werten –3 und 3 gekappt. Jedes Rautenzeichen („#“) steht für 22 Pbn, jeder Punkt für 1 bis 21 Pbn. Im rechten Teil der Abbildung sind die Texte durch zweistellige Zahlen dargestellt. Jede Zahl bezeichnet das Erprobungsset, dem der betreffende Text angehört. So steht z.B. „14“ für einen Text aus dem Erprobungsset E14. Die Buchstaben „S“ und „T“ entlang der gestrichelten Trennlinie in der Mitte definieren den Abstand vom Mittelwert „M“ der jeweiligen Verteilung der Logitwerte in Einheiten der Standardabweichung („S“ = 1 Standardabweichung, „T“ = 2 Standardabweichungen).

Anhand der Nummern der einzelnen Texte ist gut zu erkennen, dass ein Text aus E14 der schwerste, ein Text aus E06 mit Abstand der leichteste von allen Texten ist. Auch andere Texte aus E14 erweisen sich in Relation zu den Texten der übrigen Sets als schwer. Welche Sets insgesamt eher schwer, welche eher leicht sind, zeigen die für jedes Set separat berechneten Mittelwerte der im Zuge der simultanen Rasch-Analyse reskalierten (verankerten) Schwierigkeitsmaße. Den höchsten mittleren Schwierigkeitsgrad weisen die Sets E16 (M = 0.67 Logits, SD = 0.85) und E14 (M = 0.55 Logits, SD = 1.02) auf. Die beiden leichtesten Sets sind E06 (M = –0.56 Logits, SD = 0.83) und E05 (M = –0.38 Logits, SD = 0.67).[9]

Die Schätzungen für die Schwierigkeitsparameter der Texte dienten einer Einteilung der Texte nach vier Schwierigkeitsstufen (leicht, mittelschwer, schwer, sehr schwer). Diese Stufen wurden als kategoriale Attribute mit den betreffenden Texten in der Itembank gespeichert. Erforderlich war die Klassifikation der Texte nach ihrer Schwierigkeit, um bei jeder Durchführung des onDaF eine systematische Staffelung der dargebotenen Texte von leichten bis hin zu sehr schweren Texten gewährleisten zu können.

Drei Aspekte verdienen bei der grafischen Darstellung der Parameterschätzungen besondere Beachtung: (a) die Verteilung der Fähigkeitsmaße (Pbn-Logits) folgt annähernd der Normalverteilung, (b) die Verteilung der Fähigkeitsmaße und die Verteilung der (zentrierten) Schwierigkeitsmaße (Texte-Logits) stehen sich nahezu spiegelbildlich gegenüber, (c) mit Abstand am meis-

[9] Es ist keineswegs ein Zufall, dass die Sets mit höheren Nummern im Durchschnitt schwerere Texte enthalten als jene mit niedrigeren Nummern. Nach den ersten Erprobungen hatte sich in der simultanen Rasch-Analyse gezeigt, dass den besonders leistungsfähigen Pbn zu wenige angemessen schwere Texte gegenüberstanden. Ähnlich unterbesetzt war die Kategorie der leichten Texte. Spätere Sets waren gezielt gebildet worden, um diese Lücken zu schließen.

ten Texte finden sich im mittleren Schwierigkeitsbereich, d.h. innerhalb einer Standardabweichung um den Nullpunkt der Skala.

Die simultane Rasch-Analyse liefert eine Klassenseparation für die Pbn von 7.61. Danach ist die Anzahl statistisch reliabel unterscheidbarer Klassen von Pbn fast doppelt so groß wie die Anzahl von Kompetenzniveaus, die anhand der Ergebnisse im onDaF separiert werden sollen. Mit anderen Worten, die Messgenauigkeit liegt weit höher als theoretisch für den onDaF zu fordern wäre (vgl. auch Kaftandjieva, 2004). Für die Pbn-Separationsreliabilität ergibt sich mit .97 eine Schätzung, die auf ähnlich hohem Niveau liegt wie bei den Analysen der Daten aus den einzelnen Erprobungen.

Zusammengenommen sprechen diese Befunde dafür, dass die Gesamtmenge der Texte ein hohes Maß an Differenzierungsfähigkeit besitzt. Anders ausgedrückt, die meisten Texte sind für einen Großteil der Pbn weder zu leicht noch zu schwer. Die Differenzierungsfähigkeit ist dort am höchsten, wo sich die Parameterschätzungen für die Pbn am stärksten häufen.[10]

3.2. Standard-Setting mit der Prototypgruppenmethode

Auf der Basis der simultanen Rasch-Analyse wurden Cut-Scores berechnet, um zu einer Einteilung der Testpersonen in die vier angestrebten Leistungsklassen oder Kompetenzniveaus (A2, B1, B2, C1) zu gelangen. Da Erprobungen beim onDaF, wie bereits ausgeführt, fortlaufend durchgeführt und analysiert werden, sind die zu einem bestimmten Zeitpunkt ermittelten Cut-Scores stets nur als vorläufig und prinzipiell veränderbar bzw. im Lichte der jeweils vorhandenen Datenbasis revidierbar zu betrachten. Hinzu kommt, dass separate Studien zur Validierung der onDaF-Kompetenzstufen, wie z.B. die später geschilderte DIALANG-Vergleichsstudie, zu Erkenntnissen führen können, die ebenfalls eine Revision von Cut-Scores nach sich ziehen. Schließlich müssen sich Cut-Scores auch in der Testpraxis bewähren, d.h., sie müssen sich im jeweiligen Anwendungsgebiet als nützlich erweisen. Tun sie dies nicht im gewünschten Ausmaß, wäre auch hierin eine Indikation für eine Revision zu sehen.

3.2.1. Einstufung der Lernerprototypen

Das Standard-Setting folgte der oben beschriebenen Prototypgruppenmethode. Danach hatten Sprachlehrer bzw. Kursleiter gleichzeitig mit der Erprobung eines Sets von 10 Texten maximal drei Deutschlerner zu benennen, die sie als beson-

[10] Zu leichte Texte wären (bildlich gesprochen) relativ zu den Fähigkeitsschätzungen entlang der Logitskala nach unten verschoben; zu schwere Texte wären entsprechend nach oben verschoben.

ders typische Vertreter eines bestimmten Kompetenzniveaus ansahen. Sie waren gehalten, diese Benennungen nur für jene Lerner vorzunehmen, deren Sprachkenntnisse sie zum Zeitpunkt der Erprobung sehr gut einschätzen konnten. Die Einschätzung der Sprachkenntnisse sollte sich dabei an den Stufenbeschreibungen der globalen GER-Skala orientieren.

Von den 3.651 Testpersonen, die an den Erprobungen teilgenommen hatten, wurden insgesamt 878 Personen (24.0%) als typische Vertreter einer der vier Stufen genannt. Wie sich die Nennungen auf die Stufen verteilten, zeigt Tabelle 3.

Tabelle 3: Deskriptive Statistiken der Logitverteilungen von Lernerprototypen

Kategorie	***n***	**%**	***M***	***SD***
A2	176	4.8	–0.91	0.79
B1	246	6.7	–0.16	0.77
B2	279	7.6	0.57	0.78
C1	177	4.8	1.19	0.85
Gesamt	878	24.0	0.19	1.07

Anmerkung: n = Anzahl der Lernerprototypen pro Kategorie. Prozentangaben beziehen sich auf die Gesamtzahl der Testpersonen (N = 3.651). Die Logit-Mittelwerte für die Kategorien A2 bis C1 unterscheiden sich in Paarvergleichen statistisch signifikant voneinander (jeweils $p < .001$).

Am meisten Lernerprototypen wurden für die Kategorie B2 genannt, dicht gefolgt von der Kategorie B1. Auf die beiden Kategorien A2 und C1 entfielen die wenigsten Nennungen.

Die Mittelwerte der Verteilungen von Logitwerten der Lernerprototypen stiegen monoton von A2 (–0.91 Logits) bis C1 (1.19 Logits) an. Entsprechend hoch waren Einstufungen und Logits der Prototypen miteinander korreliert, $r(878) = .67$, $p < .001$ (Spearmans Rho = .68, $p < .001$).

Eine einfaktorielle Varianzanalyse wies die Logit-Mittelwerte als statistisch hochsignifikant verschieden voneinander aus: $F(3, 874) = 243.85$, $p < .001$, $\eta^2 = .46$. Einzelvergleiche (Scheffé-Test, Tukey-HSD-Test) belegten zudem, dass die Unterschiede zwischen den Mittelwerten sämtlich hochsignifikant waren (alle $p < .001$).

Diese Ergebnisse lassen den Schluss zu, dass die Einstufungen der Lernerprototypen in systematisch abgestufter Weise auf der Basis der vier vorgegebenen Kategorien erfolgt waren. Mit anderen Worten, Lerner-Einstufungen und die unabhängig von diesen Einstufungen in der simultanen Rasch-Analyse ermittelten Fähigkeitsparameter stimmten in zufrieden stellend hohem Maße überein.

Die Prototypgruppenmethode des Standard-Settings konnte sich demnach auf eine valide Datenbasis stützen.

Um die Cut-Scores festzulegen, wurden zwei verschiedene statistische Verfahren angewendet: (a) das Median-Verfahren, (b) die binäre logistische Regression.

3.2.2. Median-Verfahren

Im Median-Verfahren wird das Intervall der Überlappung zwischen den Logitverteilungen zweier benachbarter Leistungskategorien berechnet. Dieses Intervall ist definiert als der Wertebereich zwischen dem niedrigsten Logitwert der höheren Kategorie und dem höchsten Logitwert der niedrigeren Kategorie. Der jeweilige Median der Verteilung von Logitwerten im Überlappungsintervall liefert (nach entsprechender linearer Transformation) eine Schätzung des Cut-Scores.

Da die Ergebnisse des Median-Verfahrens sehr stark von einzelnen Logitwerten abhängen und damit von möglichen „Ausreißern" (d.h. von extrem abweichenden Einstufungen) beeinflusst sind, wurden vor der Bestimmung der Intervallgrenzen in jeder Verteilung die oberen und unteren 10% der Pbn eliminiert. Das heißt z.B., dass in der Verteilung der typischen B2-Lerner 28 Pbn mit den höchsten Logitwerten und die gleiche Anzahl von Pbn mit den niedrigsten Logitwerten unberücksichtigt blieben. Auf diese Weise wurde sichergestellt, dass sich die Medianberechnungen auf eine eher konsensuelle, nicht von idiosynkratischen Einschätzungen der Sprachkenntnisse einzelner Pbn verzerrte Datenbasis bezogen.[11]

Die schließlich ermittelten Mediane (*Mdn*) lauten in Einheiten der Logitskala wie folgt (in Klammern die Anzahl der Pbn im jeweiligen Intervall): *Mdn* (A2 vs. B1) = –0.58 (207), *Mdn* (B1 vs. B2) = 0.20 (287), *Mdn* (B2 vs. C1) = 0.92 (284).

Die Werte der Mediane lassen sich, da sie in Logits angegeben sind, direkt auf die Verteilung der Personenparameter von Abbildung 3 anwenden. Für Zwecke der Einteilung von Testpersonen anhand ihrer im onDaF erreichten Summenscores sind diese Cut-Scores noch durch eine geeignete lineare Transformation in die onDaF-Skala zu überführen (vgl. Linacre, 2007).

---

[11] Die auf 38 Länder aus fünf Kontinenten verteilten Kursleiter hatten kein Training in der korrekten, einheitlichen Anwendung der GER-Stufenbeschreibungen erhalten. Daher war ein gewisses Maß an Unterschiedlichkeit im Verständnis der Lernerprototypen zu vermuten. Diese Unterschiedlichkeit sollte durch den vorbereitenden Schritt der Kappung der Logitverteilungen im Sinne einer Ausreißerkontrolle (vgl. z.B. Myers & Well, 2003) in Grenzen gehalten werden.

3.2.3. Binäre logistische Regression

Das zweite Verfahren stützte sich auf das Modell der binären logistischen Regression (vgl. z.B. Backhaus, Erichson, Plinke & Weiber, 2006; Cohen, Cohen, West & Aiken, 2003; Rudolf & Müller, 2004). Dieses Regressionsmodell erlaubt die Schätzung der Wahrscheinlichkeit von Werten einer dichotomen Kriteriumsvariablen aufgrund der Kenntnis der Werte einer (oder mehrerer) Prädiktorvariablen. Im Fall der beiden Kompetenzstufen A2 und B1 hatte die Kriteriumsvariable die Werte „Zugehörigkeit zu Kategorie A2“ und „Zugehörigkeit zu Kategorie B1“. Als Prädiktorvariable dienten die Schätzungen der Personenparameter in Einheiten der Logitskala.

In der logarithmischen Schreibweise hat die binär-logistische Regressionsgleichung mit einem einzigen Prädiktor *X* die folgende Form:

$$\ln\left(\frac{p(y_i=1)}{1-p(y_i=1)}\right)=b_0+b_1x_i. \qquad (1)$$

Dabei ist $p(y_i = 1)$ die vorhergesagte Wahrscheinlichkeit der Zugehörigkeit einer Person *i* zu einer definierten Kategorie, x_i ist der Logitwert von Person *i*, b_0 ist die Regressionskonstante und b_1 der Regressionskoeffizient („ln“ steht für den natürlichen Logarithmus).

Setzt man in (1) für $p(y_i = 1)$ den Wert .50 ein und löst nach $x_i = x_c$ (d.h. nach dem Logitwert der exakt zwischen den beiden Kategorien lokalisierten Testperson) auf, so ergibt sich $x_c = -b_0/b_1$. Mit x_c ist der Cut-Score für die beiden betreffenden Kategorien in Einheiten der Logitskala gefunden (vgl. Livingston & Zieky, 1989).

In Tabelle 4 sind die Ergebnisse der verschiedenen Regressionsanalysen nach Modellgleichung (1) zusammengefasst.

Tabelle 4: Ergebnisse der logistischen Regressionsanalysen zur Bestimmung von Cut-Scores

Niveauvergleich	***n***	$\boldsymbol{b_0}$	$\boldsymbol{b_1}$	***SE***	**Fit**	$\boldsymbol{x_c}$
A2 versus B1	379	1.997**	3.031**	0.321	.56	–0.66
B1 versus B2	472	–0.441*	2.657**	0.249	.52	0.17
B2 versus C1	410	–2.271**	2.080**	0.227	.41	1.09

Anmerkung: *n* = Anzahl der Lernerprototypen in den beiden jeweils betrachteten Kategorien. b_0 = Regressionskonstante. b_1 = Regressionskoeffizient. *SE* = Standardfehler (Regressionskoeffizient). Fit = Nagelkerke-R^2-Index (gibt die Modellanpassung analog zu einem Maß der Varianzaufklärung an). x_c = Cut-Score in Einheiten der Logitskala. * $p < .01$. ** $p < .001$.

Alle Regressionskoeffizienten erweisen sich als statistisch hochsignifikant. Die Modellgüte wird durch den Nagelkerke-R^2-Index ausgedrückt. Dieser Index kann ähnlich wie ein Maß der Varianzaufklärung in der linearen Regressionsanalyse interpretiert werden. Bei den ersten beiden Vergleichen (A2 vs. B1, B1 vs. B2) bewegt sich der Index auf einem Niveau der Modellanpassung, das als „sehr gut" gelten kann; beim Vergleich zwischen B2 und C1 zeigt der Index immer noch eine „gute" Modellanpassung an (vgl. z.B. Backhaus et al., 2006). Die in der letzten Spalte von Tabelle 4 angegebenen Cut-Scores zeichnen sich demnach durch eine hohe Verlässlichkeit aus. Im Vergleich mit den Cut-Scores, die nach der Median-Methode resultierten, sind keine großen Verschiebungen festzustellen. Allenfalls sind die mittleren Kategorien (B1 und B2) nach der logistischen Regressionsmethode etwas breiter definiert.

Die bisherigen Analysen haben gezeigt, wie sich Cut-Scores für die Trennung zwischen den Niveaustufen A2 und B1, B1 und B2 sowie B2 und C1 empirisch definieren lassen. Wie bereits ausgeführt, ist die onDaF-Stufe C1 nach oben offen, d.h., sie umfasst alle Kompetenzniveaus oberhalb von B2 („C1" ist eine Kurzform für „C1 oder höher"). Ein oberer Cut-Score (zur Abgrenzung von C2) wird daher nicht benötigt. Anders verhält es sich bei der onDaF-Stufe A2. Um diese Stufe eindeutig zu bestimmen, ist es erforderlich, auch einen „unteren" A2-Cut-Score, d.h. einen Cut-Score zwischen A2 und den darunter liegenden Kompetenzbereichen, festzulegen. Dieser Cut-Score wurde auf der Basis der Ergebnisse eines Vergleichs zwischen den onDaF-Stufen und den von DIALANG geleisteten Einstufungen ermittelt. Der nächste Abschnitt geht hierauf näher ein.

4. Validierung

4.1. Konstruktvalidität von C-Tests

Es liegt eine Fülle empirischer Untersuchungen vor, die sich mit der Frage der Validität von C-Tests befassen. Diese Frage lautet: Messen C-Tests tatsächlich das, was sie messen sollen – allgemeine Sprachkompetenz in einer Fremdsprache oder in der Muttersprache?

Neuere Übersichtsdarstellungen finden sich z.B. bei Eckes & Grotjahn (2006a), Grotjahn, Klein-Braley & Raatz (2002) sowie Sigott (2004). Am häufigsten wurde die Kriteriumsvalidität englischer C-Tests untersucht. Als Kriterien dienten dabei international renommierte Tests wie der „Test of English as a Foreign Language" (TOEFL) oder der „Test of English for International Communication" (TOEIC). Die Korrelationen mit dem TOEFL-Gesamtscore lagen mindestens bei .55, die mit dem TOEIC-Gesamtscore mindestens bei .62 (vgl.

Eckes & Grotjahn, 2006a, Tab. 1). Aber auch auf der Ebene der einzelnen Sprachfertigkeiten wurden substanzielle Korrelationen (zumeist um .50 oder darüber) berichtet.

Von besonderem Interesse sind die Korrelationen eines deutschen C-Tests mit den Subtests des TestDaF (Eckes & Grotjahn, 2006a). Dieser C-Test bestand aus vier Texten, die in gleicher Weise aufgebaut waren wie die onDaF-Texte. Die Einzelkorrelationen beliefen sich auf .56 für Leseverstehen, .62 für Hörverstehen, .58 für Schriftlichen Ausdruck und .65 für Mündlichen Ausdruck (alle Korrelationen $p < .01$, $N = 470$). Für die multiple Korrelation resultierte ein statistisch hochsignifikanter Wert von $R = .76$ ($p < .01$, $N = 470$).

Um die Zusammenhänge zwischen diesem C-Test und den TestDaF-Subtests genauer zu analysieren, führten Eckes und Grotjahn (2006a) eine Reihe von konfirmatorischen Faktorenanalysen durch. Es zeigte sich, dass ein einfaktorielles Modell mit dem Faktor „allgemeine Sprachkompetenz“ die Zusammenhänge zwischen dem C-Test und den TestDaF-Subtests zufrieden stellend abbildete. Dabei hatte der C-Test mit .83 eine höhere Ladung auf diesem Faktor als jeder der fertigkeitsspezifischen Tests. Abbildung 4 veranschaulicht die faktorielle Struktur.

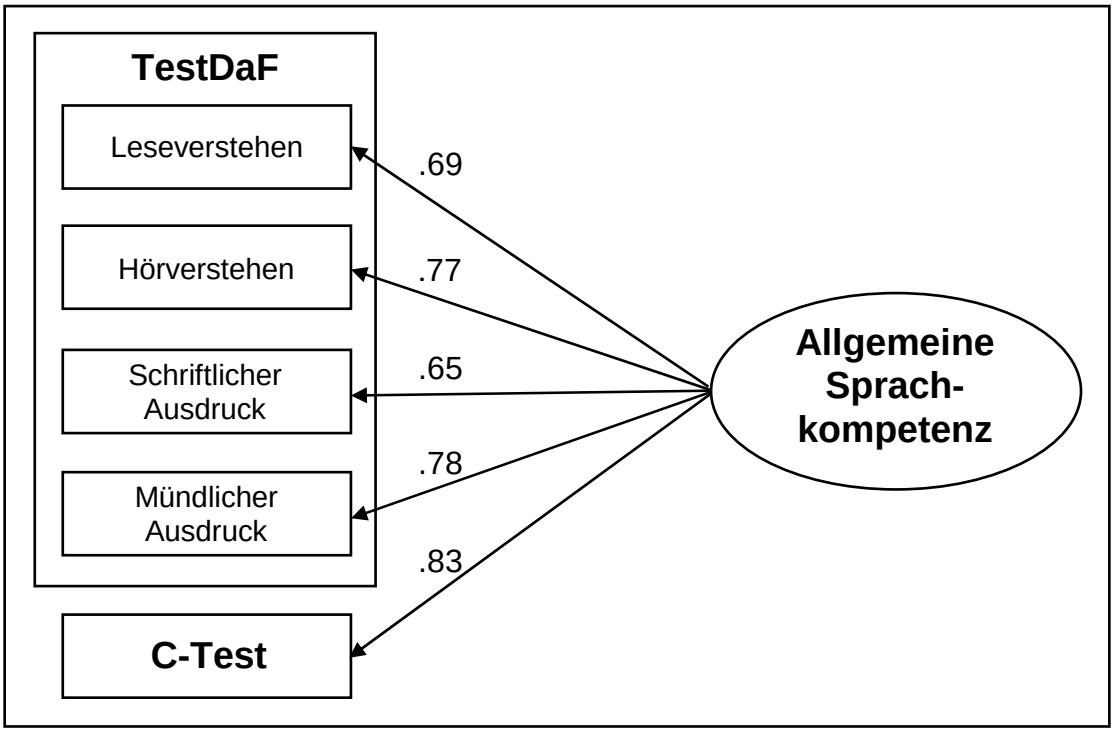

Abbildung 4: TestDaF-Subtests und C-Test in Relation zum Faktor der allgemeinen Sprachkompetenz. Der C-Test hat mit .83 die höchste Ladung auf diesem Faktor (nach Eckes & Grotjahn, 2006a, S. 301)

Die bemerkenswert engen Beziehungen zwischen dem C-Test und den produktiven Fertigkeiten des Schreibens und Sprechens kamen auch darin zum Ausdruck, dass von den ebenfalls untersuchten zweifaktoriellen Modellen jenes die beste Anpassung aufwies, bei dem der C-Test dem Produktionsfaktor (zusammen mit Schreiben und Sprechen) und nicht dem Rezeptionsfaktor (mit Leseverstehen und Hörverstehen) zugeordnet war.[12]

Andere wichtige Kriteriumsvariablen betreffen Tests der grammatikalischen und der lexikalischen Kompetenz. Wenn C-Tests allgemeine Sprachkompetenz messen, dann sollten nicht nur substanzielle Korrelationen mit den rezeptiven und produktiven Sprachfertigkeiten, sondern auch mit Kenntnissen der Grammatik und des Wortschatzes zu beobachten sein. Nach Singleton (1999) erfassen C-Tests lexiko-grammatikalische Sprachkenntnisse und korrelieren deshalb mit so vielen anderen Tests der allgemeinen Sprachkompetenz, weil das mentale Lexikon den Kernbereich verschiedener Formen des Gebrauchs von Sprache bildet. Ähnlich schreibt Read (2000) der lexikalischen Kompetenz eine wesentliche Rolle bei der Verarbeitung von Texten eines C-Tests zu. Tatsächlich liegen die in der Forschungsliteratur berichteten Korrelationen mit Grammatik- und Wortschatztests in der Mehrzahl der Fälle um .70 bis .80. Diese und andere Befunde belegen, dass C-Tests das messen, was sie messen sollen.

Fast alle Studien verwendeten allerdings C-Tests in der Papierversion. Es stellt sich daher die Frage, wie die Konstruktvalidität von C-Tests beschaffen ist, wenn die Testdurchführung computergestützt erfolgt. Diese Frage ist ganz offenkundig von großer Bedeutung für die Interpretation von Ergebnissen, die Testpersonen im onDaF erzielen. Grundsätzlicher betrachtet handelt es sich um die Frage der Äquivalenz zwischen Computer- und Papierversion eines Tests (vgl. Choi, Kim & Boo, 2003; International Test Commission, 2006; Klinck, 2002, 2006).

Computergestützte C-Tests gibt es seit Mitte der 1990er Jahre (Germann & Grotjahn, 1994; Koller & Zahn, 1996). Die Äquivalenz von Computer- und Papier-C-Tests wurde erstmals von Bisping & Raatz (2002) untersucht. In dieser Studie ergab sich eine hohe psychometrische Äquivalenz beider Testversionen, d.h., die Mittelwerte und Standardabweichungen der betreffenden Scoreverteilungen waren nahezu identisch, die Scores verteilten sich in beiden Fällen nor-

[12] Auch wenn wiederholt enge Zusammenhänge zwischen C-Tests und fertigkeitsbezogenen Sprachtests beobachtet wurden, ist keineswegs ausgeschlossen, dass es (bei hoher allgemeiner Sprachkompetenz) beträchtliche inter- oder intraindividuelle Unterschiede in einzelnen Sprachfertigkeiten gibt. Das Fertigkeitsprofil einer Testperson hängt von vielen Faktoren ab, die hauptsächlich in der individuellen Lerngeschichte zu suchen sind. C-Tests lassen hierüber nur allgemeine, fertigkeitsübergreifende Schlussfolgerungen zu.

mal und die Reliabilität der Computerversion (.85) lag ähnlich hoch wie die Reliabilität der Papierversion (.84). Ferner zeigte sich, dass Computerangst bzw. Computererfahrung ebenso wenig die Testergebnisse beeinflussten wie das Geschlecht der Testpersonen oder das subjektive Erleben des Tests.

In einer Nachfolgeuntersuchung verglich Bisping (2006) beide C-Test-Versionen unter Verwendung eines Multitrait-Multimethod-Designs. Die Ergebnisse waren wieder eindeutig: Computer- und Papierversion zeigten eine hohe konvergente Validität ($r = .79$, $p < .01$). Zugleich besaß die Computerversion diskriminante Validität, d.h., sie korrelierte mit einem ebenfalls am Computer durchgeführten Konzentrationstest signifikant niedriger als mit der Papierversion des C-Tests. Der Autor kam zu dem Schluss, dass computerisierte C-Tests nicht weniger valide seien als traditionelle C-Tests (vgl. auch Reichert, Keller & Martin, im vorliegenden Band).

4.2. Externe Validierung: DIALANG als Kriteriumstest

Die Befunde langjähriger C-Test-Forschung mit zum Teil sehr unterschiedlichen methodischen Ansätzen lassen keinen Zweifel daran, dass C-Tests ein zuverlässiges Instrument zur Messung der allgemeinen Sprachkompetenz sind. Empirisch nachgewiesene enge Zusammenhänge mit den vier Sprachfertigkeiten sowie mit grammatikalischen und lexikalischen Kenntnissen belegen, dass sich C-Tests nicht nur durch hohe Reliabilität, sondern auch durch hohe Validität auszeichnen. Zudem unterstreicht die Invarianz von C-Tests gegenüber dem Medium der Testdarbietung (computergestützte Darbietung oder traditionelle Papierversion) ihre Robustheit.

Was den onDaF betrifft, so erhärtet diese Befundlage zwar das grundlegende Validitätsargument. Dennoch bleiben zwei Fragen zu beantworten: (a) Wie valide sind die bei der internetgestützten Durchführung des onDaF erzielten Testscores? (b) Wie valide sind die vom onDaF vorgenommenen Einstufungen der Sprachkompetenz von Testpersonen?

Die umfangreichen Skalenanalysen im Rahmen der Erprobungen von Texten für den onDaF erlaubten den Aufbau einer kalibrierten Itembank mit Texten hoher psychometrischer Qualität. Da jedoch bei jeder onDaF-Durchführung Texte aus der Itembank neu ausgewählt und zur Bearbeitung auf dem Bildschirm angezeigt werden (siehe hierzu den folgenden Abschnitt), ist es erforderlich, unabhängige Evidenz für die Validität der von den Testpersonen erzielten Ergebnisse (d.h. Summenscores und Einstufungen) vorzulegen.

Hierzu führte ich eine so genannte **externe Validierung** durch (vgl. Council of Europe, 2003; Figueras, North, Takala, Verhelst & van Avermaet, 2005; North, 2004). Als externes Validitätskriterium fungierte der Deutschtest aus dem

Online-Testsystem DIALANG (Alderson, 2005; Alderson & Huhta, 2005; Huhta et al., 2002).[13]

Im Folgenden gehe ich zunächst auf den Standardmodus der Testdarbietung beim onDaF ein und bespreche anschließend allgemeine Merkmale von DIALANG. Die weiteren Abschnitte befassen sich mit dem Aufbau der Validierungsstudie und ihren Ergebnissen.

4.2.1. Darbietung des onDaF

Tests lassen sich auch nach der Art und Weise unterscheiden, wie sie Testpersonen dargeboten werden (vgl. Drasgow, Luecht & Bennett, 2006; Folk & Smith, 2002). Weit verbreitet sind die folgenden Varianten: (a) „fixierte Tests" oder Tests mit festem Format (alle Testpersonen erhalten denselben Test, d.h. dieselben Items in derselben Reihenfolge), (b) „Linear-on-the-Fly"-Tests (jede Testperson erhält eine andere, neu zusammengestellte Menge äquivalenter Items; „LOFT-Methode" der Testdarbietung), und (c) computeradaptive Tests (CATs; jede Testperson erhält die Items, die ihrem Fähigkeitsniveau bestmöglich entsprechen).

Beim onDaF kommt eine Variante der LOFT-Methode zum Einsatz.[14] Diese Variante sieht folgenden Ablauf vor. Sobald eine Testperson den Test startet, öffnet sich das onDaF-Testfenster und der erste Text wird zur Bearbeitung auf dem Bildschirm dargeboten. Es stehen während der Textbearbeitung keinerlei Browserfunktionen zur Verfügung. Die Bearbeitungszeit pro Text beträgt maximal fünf Minuten. Daraus ergibt sich eine maximale Bearbeitungsdauer des gesamten Tests von 40 Minuten.

Nach genau fünf Minuten Bearbeitungszeit pro Text werden die Teilnehmer automatisch zum nächsten Text weitergeleitet. Die letzten 60 Sekunden vor der Weiterleitung werden angezeigt. Hat ein Teilnehmer die Bearbeitung eines Tex-

---

[13] In der Terminologie des GER-Manuals des Europarats (Council of Europe, 2003, S. 109) handelt es sich im vorliegenden Kontext beim DIALANG-Deutschtest um einen "Ankertest", da er am GER „kalibriert", d.h. in direktem Bezug zum GER entwickelt wurde. Abweichend hiervon reserviere ich den Begriff „Ankertest" für einen Satz gemeinsamer Items im Rahmen eines Equating-Designs. Der Deutschtest von DIALANG ist genauer als Kriteriumstest zu charakterisieren, an dem der onDaF validiert werden soll. Diese Validierung lässt sich auch als die Herstellung einer indirekten Beziehung zum GER („indirektes Linking") auffassen.

[14] Im Falle des (ohnehin kurzen und messgenauen) onDaF bietet die LOFT-Methode gegenüber einem CAT die folgenden Vorteile für die Testpraxis: höhere Robustheit gegenüber Schwankungen in den Bedingungen der Testdurchführung, geringere Anforderungen an den Umfang der Itembank und bessere Ausschöpfung der thematischen Vielfalt der Texte pro Testung. Der letzte Punkt ist insbesondere für die Untersuchung von Lernfortschritten von Bedeutung (vgl. Lehmann, 2003).

tes vor Ablauf der fünf Minuten abgeschlossen, kann er durch Mausklick zum nächsten Text gelangen. Ein Zurückgehen zu früher bearbeiteten Texten ist nicht möglich. Das onDaF-Testfenster lässt sich erst nach der Bearbeitung des letzten Textes und der Rückmeldung der Testergebnisse schließen. Es werden keine Teilnehmerdaten oder Testergebnisse auf dem Teilnehmerrechner gespeichert.

Die Zufallsauswahl der Texte aus der kalibrierten Itembank unterliegt einer zweifachen Einschränkung. So muss der jeweilige Text (a) die geforderte Schwierigkeitsstufe aufweisen und (b) zu einer Themenkategorie gehören, die in keinem der zuvor ausgewählten Texte aufgetreten ist. Die Darbietung der Texte erfolgt grundsätzlich nach ansteigender Schwierigkeit, d.h., zuerst werden zwei Texte aus der niedrigsten Schwierigkeitsstufe dargeboten, dann erscheinen zwei Texte aus der nächst höheren Schwierigkeitsstufe usw. Auf jeder Stufe sind genau zwei Texte zu bearbeiten. Dies sichert bei geringer maximaler Bearbeitungsdauer des gesamten Tests eine hinreichend hohe Messgenauigkeit. Voruntersuchungen hatten gezeigt, dass etwa bei drei Texten pro Stufe die Reliabilität nur unwesentlich höher lag (bei einer maximalen Bearbeitungsdauer des gesamten Tests von 60 Minuten). Innerhalb einer Stufe wird stets zuerst der etwas leichtere, dann der etwas schwierigere Text dargeboten.[15]

4.2.2. DIALANG

Bei DIALANG handelt es sich um ein internetgestütztes Sprachtestsystem, das Tests in 14 europäischen Sprachen anbietet und direkt auf den GER bezogen ist. Hauptziel von DIALANG ist die Diagnose von Kenntnissen und Fertigkeiten eines Lerners in der Zielsprache. Testpersonen erhalten eine detaillierte Rückmeldung zu ihren Antworten unmittelbar nach der Bearbeitung einer Aufgabe bzw. eines Prüfungsteils. Auf eine Zertifizierung wird ausdrücklich verzichtet. Eine umfassende Darstellung von Konzeption und Konstruktion dieses Testsystems gibt Alderson (2005). Eine kurz gefasste Übersicht findet sich bei Alderson & Huhta (2005). Ausgewählte Aspekte von DIALANG stellen Fischer (2000), Mackiewicz (2001) und von der Handt (2001) dar (vgl. auch Europarat, 2001, Anhang C).

DIALANG besteht aus fünf Prüfungsteilen: Hörverstehen, Schreiben, Leseverstehen, Grammatik (Strukturen) und Wortschatz. Jeder Prüfungsteil enthält 30 Items. Die in jedem Prüfungsteil dargebotenen Items sind einem von drei

---

[15] Ein mathematischer Algorithmus stellt zudem sicher, dass die bei verschiedenen Testdarbietungen pro Schwierigkeitsstufe ausgewählten Texte im Mittel gleich schwierig sind. Dieser Algorithmus stand allerdings zum Zeitpunkt der Validierungsstudie nicht zur Verfügung.

Schwierigkeitsniveaus (leicht, mittel, schwer) zugeordnet. Das jeweilige Schwierigkeitsniveau wird anhand eines vorgeschalteten Einstufungstests ("Vocabulary Size Placement Test", kurz VSPT) und anhand von Selbsteinschätzungen in den Fertigkeiten Hörverstehen, Schreiben und Leseverstehen ermittelt ("Test-Level Adaptivity"; Alderson, 2005).

Der VSPT enthält eine Liste von 75 Wörtern (allesamt Verben), davon stammen 50 Wörter aus der Zielsprache, die restlichen 25 sind Pseudowörter. Aufgabe der Testpersonen ist es, bei jedem Wort anzugeben, ob es sich um ein richtiges Wort oder um ein Pseudowort handelt.

DIALANG nutzt insgesamt vier verschiedene Item-Formate: Multiple-Choice, Drop-down-Menü, Texteingabe und Kurzantwort. Die hauptsächlich verwendeten Formate pro Prüfungsteil sind: Im Hörverstehen Multiple-Choice, im Schreiben Drop-down und Texteingabe, im Leseverstehen Multiple-Choice und Drop-down, in der Grammatik Multiple-Choice und Texteingabe sowie im Wortschatz Multiple-Choice, Drop-down und Texteingabe.

In jedem Prüfungsteil wird die Leistung der Testpersonen auf der globalen GER-Skala von A1 bis C2 eingestuft. Aufgrund seiner sprachdiagnostischen Ausrichtung gibt DIALANG keine Einstufung, die die Leistung über alle Prüfungsteile hinweg zusammenfasst. Das Testsystem ist kostenlos und steht auf jedem Rechner mit Internetzugang zur Verfügung (www.dialang.org).

DIALANG wurde in den Jahren zwischen 1996 und 2004 mit Mitteln der Europäischen Kommission und ca. 25 weiterer europäischer Institutionen (überwiegend Hochschulen) entwickelt. Erstmals orientierte sich die Entwicklung eines Sprachtests von Beginn an unmittelbar am GER (Alderson, 2005; vgl. auch Little, 2006). Die Erprobungen der Items konnten sich allerdings nur in vier Sprachen auf eine ausreichende empirische Datenbasis stützen: Deutsch (N = 707), Englisch (N = 2.265), Französisch (N = 702) und Spanisch (N = 680).

4.2.3. Zielsetzung

Beim Vergleich von onDaF- mit DIALANG-Einstufungen ist zu berücksichtigen, dass der onDaF allgemeine Sprachkompetenz misst, DIALANG im Unterschied hierzu drei Sprachfertigkeiten (Hörverstehen, Schreiben, Leseverstehen) und zwei Sprachbereiche (Grammatik, Wortschatz) differenziert erfasst. Auch wenn die Testkonstrukte divergieren, so besitzt doch DIALANG eine Reihe von Eigenschaften, die dieses System als Kriteriumstest für onDaF geeignet erscheinen lassen. Die im Rahmen der Validierungsstudie relevanten Vorteile lauten: DIALANG ist (a) ein internetgestützter Test, (b) direkt auf den GER bezogen und (c) als Deutschtest mit kalibrierten Items verfügbar.

Im Einzelnen verbinden sich mit der Validierungsstudie die nachfolgend aufgeführten Erwartungen bzw. Ziele. Dabei beziehen sich die Punkte (1) bis (3) auf die onDaF- und DIALANG-Testscores, die Punkte (4) und (5) auf die Einstufungen der jeweiligen Teilnehmerleistungen.

(1) Die auf der Basis der Testscores ermittelten Korrelationen zwischen onDaF und den DIALANG-Prüfungsteilen Hörverstehen, Schreiben, Leseverstehen, Grammatik und Wortschatz sollten in ähnlicher Höhe liegen wie die aus der Literatur bekannten Korrelationen zwischen C-Tests und den entsprechenden fertigkeits- bzw. bereichsspezifischen Tests, d.h. etwa zwischen .50 und .70.

(2) Die einzelnen DIALANG-Prüfungsteile sollten sich in einer linearen Regressionsanalyse als signifikante Prädiktoren der onDaF-Testscores erweisen. Die multiple Korrelation sollte statistisch hochsignifikant und substanziell sein, d.h. mindestens .70 betragen.

(3) Ähnlich wie in der konfirmatorischen Faktorenanalyse von Eckes & Grotjahn (2006a) sollten der onDaF und die fünf DIALANG-Prüfungsteile auf einem gemeinsamen Faktor hoch laden.

(4) Die aus dem Standard-Setting abgeleiteten Cut-Scores sollten eine statistisch signifikante Übereinstimmung zwischen den onDaF- und den DIALANG-Einstufungen ergeben.

(5) Im Hinblick auf eine mögliche Verbesserung der Übereinstimmungsraten sind die Cut-Scores systematisch zu revidieren. Die Auswirkungen jeder einzelnen Revision auf die Übereinstimmungsraten sind zu untersuchen.

4.2.4. Teilnehmer

An der Validierungsstudie nahmen insgesamt 223 Personen teil, darunter 140 Frauen (62.8%) und 83 Männer (37.2%). Das Alter von rund 86% der Teilnehmer lag zwischen 18 und 27 Jahren (M = 24.29, SD = 9.71).

Die Teilnehmer stammten aus 47 Ländern. Am häufigsten vertreten waren die folgenden sechs Herkunftsländer (in Klammer die Teilnehmerzahl): Vietnam (40), Russische Föderation (28), Kirgistan (17), Volksrepublik China (12), Mexiko (12), Bulgarien (11). Die Pbn verteilten sich auf 34 Studienfächer bzw. Studienfelder, darunter Wirtschaftswissenschaften (36), Germanistik (34) und Übersetzen/Dolmetschen (12); zur Schule gingen noch 34 Pbn.

Zum Zeitpunkt der Datenerhebung besuchten die Pbn Sprachkurse an Testzentren des TestDaF-Instituts in neun Ländern, darunter sieben Testzentren in Deutschland, je ein Testzentrum in Brasilien, Bulgarien, Kirgistan, Luxemburg, Mexiko, Spanien, Tadschikistan und Vietnam. Alle Pbn nahmen freiwillig teil. Der Großteil der Pbn erhielt für die Teilnahme ein Entgelt in Höhe von 10 Euro.

Alle Testzentren erhielten eine Aufwandsentschädigung in Höhe von ebenfalls 10 Euro pro Teilnehmer.

4.2.5. Durchführung und Auswertung

Die Teilnehmer bearbeiteten zuerst den onDaF (Testdauer: maximal 40 Minuten) und anschließend den DIALANG-Deutschtest (Testdauer: ca. 3 bis 4 Stunden; die Dauer war abhängig von der individuellen Bearbeitungsgeschwindigkeit). Zwischen onDaF und DIALANG lag eine Pause von mindestens 30 Minuten.

Im Falle des onDaF waren nacheinander acht Texte aufsteigender Schwierigkeit zu bearbeiten (je zwei Texte aus den vier Schwierigkeitsstufen). Alle Texte wurden nach der oben beschriebenen LOFT-Methode der onDaF-Itembank entnommen.

Die Auswertung der Teilnehmerantworten erfolgte vollkommen automatisch und daher mit einem Höchstmaß an Objektivität. Grundlage hierfür waren die auf dem onDaF-Server zu jedem Text gespeicherten korrekten Ergänzungen. Als „korrekt“ waren (wie in der Erprobungsphase) ausschließlich orthografisch richtige Originale und orthografisch richtige Varianten definiert. Für jede korrekte Ergänzung wurde genau ein Punkt vergeben. Jede inkorrekte oder fehlende Ergänzung wurde mit 0 Punkten bewertet. Dies ergab ein Maximum von 160 erreichbaren Punkten.

Als einzige Abweichung vom Standardverfahren einer onDaF-Durchführung (siehe Abschnitt 4.2.1) gab es keine Rückmeldung der Ergebnisse an die Teilnehmer. Damit sollte eine mögliche Beeinflussung der Bearbeitung von Items im sich anschließenden DIALANG-Test ausgeschlossen werden. Die Ergebnisse aller Teilnehmer (Punktzahl, onDaF-Einstufung) wurden automatisch auf dem onDaF-Server gespeichert.

In DIALANG durchliefen alle Teilnehmer die Prüfungsteile in derselben Reihenfolge (einschließlich der jeweiligen Selbsteinschätzungen). Zuerst legten sie den Einstufungstest VSPT ab. Danach bearbeiteten sie die Items aus den Prüfungsteilen Hörverstehen, Schreiben, Leseverstehen, Grammatik und Wortschatz. Jeder Teilnehmer erhielt einen einseitigen Ergebnisbogen, in den die jeweils erzielten Ergebnisse einzutragen waren. Die Ergebnisse betrafen beim VSPT die online mitgeteilte Punktzahl, bei Hörverstehen, Schreiben und Leseverstehen die GER-Stufe, die Punktzahl und die vom System ermittelte Stufe der Selbsteinschätzung, bei Grammatik und Wortschatz die GER-Stufe und die Punktzahl.

Da die pro Prüfungsteil erreichte Punktzahl im DIALANG-System nicht mitgeteilt wird, hatten die Teilnehmer nach Abschluss eines Subtests im Menü den

Schritt „Prüfen Sie Ihre Antworten“ zu wählen und die richtigen, grün markierten Antworten zu zählen. Um eine für alle Teilnehmer identische Bearbeitungsweise zu gewährleisten, wurde jedem Teilnehmer eine Schritt-für-Schritt-Anleitung ausgehändigt. Die Sprache der Testdurchführung war in allen Fällen Deutsch.

Die statistischen Analysen bezogen sich einmal auf die Testscores (d.h. die Punktzahlen im onDaF und den DIALANG-Prüfungsteilen, einschl. VSPT), zum anderen auf die in beiden Tests erzielten Einstufungen. Da DIALANG von A1 bis C2 einstuft, onDaF aber nur von A2 bis C1, wurden bei DIALANG die Stufen C1 und C2 zu „C1 (oder höher)“ zusammengefasst; außerdem wurden in beiden Testverfahren Leistungen, die niedriger als A2 lagen, als „unter A2“ kategorisiert.

Da es bei der Durchführung von DIALANG zu einer Reihe unerwarteter technischer Probleme gekommen war, fiel die Anzahl der auswertbaren Daten niedriger aus als beim onDaF. Hiervon war insbesondere der Prüfungsteil Hörverstehen betroffen.

4.2.6. Ergebnisse

Tabelle 5 gibt die paarweisen Korrelationen zwischen onDaF und den in DIALANG enthaltenen Subtests (einschl. VSPT), berechnet auf der Grundlage der Testscores, wieder.

Tabelle 5: Korrelationen zwischen onDaF und DIALANG-Subtests

Test	**VSPT**	**HV**	**SN**	**LV**	**GR**	**WO**	***n***	***M***	***SD***
onDaF	.56*	.51*	.63*	.24*	.67*	.59*	223	81.31	25.97
VSPT	–	.27*	.40*	.08	.50*	.38*	195	336.89	251.56
HV		–	.59*	.35*	.45*	.44*	171	21.11	4.61
SN			–	.39*	.62*	.47*	219	18.82	3.94
LV				–	.31*	.30*	221	18.30	4.85
GR					–	.62*	219	18.12	5.54
WO						–	219	19.88	3.98

Anmerkung: VSPT = Vocabulary Size Placement Test (DIALANG-Einstufungstest). HV = Hörverstehen. SN = Schreiben. LV = Leseverstehen. GR = Grammatik. WO = Wortschatz. Die relativ niedrige Teilnehmerzahl im Hörverstehen geht auf technische Probleme während der Durchführung dieses Prüfungsteils zurück. Alle Korrelationen sind Produkt-Moment-Korrelationen berechnet auf Score-Basis: onDaF-Scores von 0 – 160, VSPT-Scores von 0 – 1000, HV-, SN-, LV-, GR- und WO-Scores von 0 – 30. * $p < .01$.

Die Korrelationen zwischen onDaF und den DIALANG-Subtests liegen in einer Größenordnung, die nach den bisherigen Ergebnissen der C-Test-Forschung zu erwarten waren – mit einer Ausnahme: onDaF korreliert mit Leseverstehen nur zu .24. Diese Korrelation ist zwar statistisch signifikant, liegt aber deutlich unter dem Niveau der üblicherweise berichteten Korrelationen (vgl. Eckes & Grotjahn, 2006a). Dass dies eine Besonderheit des DIALANG-Subtests Leseverstehen ist, zeigen die ebenfalls niedrigen Korrelationen von Leseverstehen mit den anderen DIALANG-Subtests (mit Werten zwischen .30 und .39).

Bemerkenswert ist auch, dass der DIALANG-eigene Einstufungstest VSPT durchweg niedriger mit den DIALANG-Subtests korreliert als der onDaF. Die Differenzen zwischen den entsprechenden onDaF- und VSPT-Korrelationen sind für alle fünf Subtests statistisch hochsignifikant. Besonders auffällig sind die Korrelationsunterschiede im Falle der Subtests Hörverstehen, Schreiben und Wortschatz. Die VSPT-Korrelationen liegen zudem deutlich niedriger als jene, die Alderson (2005, S. 87) für den Zusammenhang zwischen VSPT-Scores (skaliert nach der numerischen Meara-Methode) und den Subtest-Scores mitteilte. Dabei ist zu berücksichtigen, dass in der vorliegenden Analyse die VSPT-Ergebnisse (zusammen mit den Selbsteinschätzungen) in die Bestimmung des Schwierigkeitsniveaus der nachfolgenden Subtest-Items eingeflossen sind. Dies hätte eine Erhöhung der VSPT-Korrelationen mit den Subtests zur Folge haben müssen.

Beim Subtest Schreiben ist noch darauf hinzuweisen, dass es sich um einen indirekten Test der Fähigkeit zum schriftlichen Ausdruck handelt. Nach Alderson (2005) ist Schreiben, wie es in DIALANG gemessen wird, eng gekoppelt an grammatikalische und lexikalische Kenntnisse. Es verwundert daher nicht, dass der onDaF mit Schreiben ähnlich hoch korreliert wie mit den Subtests Grammatik und Wortschatz.

Eine multiple Regressionsanalyse mit den DIALANG-Subtests als Prädiktoren lieferte die in Tabelle 6 dargestellten Ergebnisse (n = 167; listenweiser Ausschluss von Pbn mit fehlenden Werten). Den größten Beitrag zur Vorhersage der Scores im onDaF leisten Schreiben und Grammatik. Aber auch Hörverstehen und Wortschatz haben Vorhersagekraft. Lediglich Leseverstehen erweist sich nicht als signifikanter Prädiktor.

Die multiple Korrelation liegt mit .76 (p < .01) genauso hoch wie im Falle der Gegenüberstellung von C-Test und TestDaF (Eckes & Grotjahn, 2006a). Da die Toleranzwerte als Maße der Multikollinearität sämtlich über der konventionellen Grenze von .10 liegen (vgl. z.B. Cohen et al., 2003), können die Schätzungen der Regressionskoeffizienten als stabil gelten. Insgesamt unterstreichen die reg-

ressionsanalytischen Ergebnisse die substanziellen Beziehungen zwischen onDaF und den Subtests aus DIALANG.

Tabelle 6: DIALANG-Subtests als Prädiktoren für den onDaF

DIALANG-Subtest	*n*	Korrelation	ß-Wert	Toleranz
Hörverstehen	171	.51**	.16*	.55
Schreiben	219	.63**	.31**	.49
Leseverstehen	221	.24**	.09	.64
Grammatik	219	.67**	.26**	.62
Wortschatz	219	.59**	.15*	.70

Anmerkung: Die multiple Korrelation beträgt .76 ($p < .01$). β = standardisierter Regressionskoeffizient. Toleranz = Kollinearitätsstatistik (sollte größer .10 sein). * $p < .05$. ** $p < .01$.

Eine ergänzende Sicht auf die Gemeinsamkeiten von onDaF und DIALANG eröffnet die exploratorische Faktorenanalyse. Der erste nach der Hauptkomponentenmethode extrahierte Faktor klärt über die Hälfte der Varianz auf (Eigenwert = 3.25, Varianzanteil = 54%), alle anderen Faktoren haben Eigenwerte kleiner 1. Üblicherweise werden Faktoren, die vier oder mehr Ladungen mit einem Absolutbetrag größer .60 aufweisen, als reliabel eingeschätzt (vgl. z.B. Stevens, 2002). Bis auf den Subtest Leseverstehen (.44) liegen die Ladungen deutlich über dieser Grenze: .80 (onDaF), .76 (Hörverstehen), .81 (Schreiben), .80 (Grammatik) und .74 (Wortschatz). Es liegt nahe, diesen Faktor ähnlich wie in der konfirmatorischen Analyse von Eckes & Grotjahn (2006a) als Faktor der allgemeinen Sprachkompetenz zu interpretieren.

Anhand der Cut-Scores, die sich nach den in Abschnitt 3.2 beschriebenen Verfahren ergeben hatten, wurden die Teilnehmerleistungen im onDaF einer der vier Kompetenzstufen A2 bis C1 zugeordnet. Das Niveau C1 umfasste, wie schon ausgeführt, alle Leistungen, die oberhalb B2 angesiedelt waren. Für Leistungen, die das Niveau A2 nicht erreicht hatten, wurde ein provisorischer Cut-Score, geschätzt auf der Basis der Pbn-Logitwerte, verwendet. Im Falle der DIALANG-Subtests wurden die im Ergebnisbogen eingetragenen GER-Stufen zugrunde gelegt. Tabelle 7 zeigt, wie sich die Teilnehmerleistungen auf die verschiedenen Stufen verteilen.

Tabelle 7: Prozentuale Häufigkeiten der Stufen beim onDaF (ursprüngliche Cut-Scores) und den DIALANG-Subtests

		DIALANG-Subtest				
Stufe	**onDaF**	**HV**	**SN**	**LV**	**GR**	**WO**
unter A2	4.5	4.1	5.9	9.5	1.4	0.5
A2	14.8	21.6	27.9	39.8	23.7	7.8
B1	51.6	25.1	43.4	33.9	40.2	45.7
B2	24.7	26.9	19.6	14.9	33.3	37.4
C1 (od. höher)	4.5	22.2	3.2	1.8	1.4	8.7

Anmerkung: Den onDaF legten 223 Pbn ab. HV = Hörverstehen (n = 171). SN = Schreiben (n = 219). LV = Leseverstehen (n = 221). GR = Grammatik (n = 219). WO = Wortschatz (n = 219).

Beim onDaF ist B1 die mit Abstand am häufigsten vertretene Stufe. Etwas mehr als die Hälfte der Pbn fällt in diese Kategorie. B1 ist auch in den DIALANG-Subtests Schreiben, Grammatik und Wortschatz die häufigste Stufe. Im Hörverstehen sind die Stufen A2 bis C1 (oder höher) mit Werten zwischen 22% und 27% etwa gleich häufig. Beim Leseverstehen fällt auf, dass die unteren beiden Stufen mit zusammen rund 49% sehr viel häufiger vorkommen als bei den anderen Subtests. Die Korrelationen zwischen den onDaF- und den DIALANG-Stufen (berechnet nach Kendalls Tau-b) spiegeln diese Relationen wider: .49 mit Hörverstehen, .62 mit Schreiben, .51 mit Leseverstehen, .58 mit Grammatik und .53 mit Wortschatz (alle Korrelationen signifikant mit $p < .01$).

Die in Tabelle 7 aufgeführten Prozentwerte liefern erste Anhaltspunkte dafür, dass die onDaF-Stufe A2 zu eng und die onDaF-Stufe B1 zu weit gefasst sein könnten. Genauere Einblicke geben Klassifikationstabellen, in denen die onDaF- und die DIALANG-Einstufungen einander gegenübergestellt sind. Aus Raumgründen erläutere ich hier nur das allgemeine Vorgehen am Beispiel des Subtests Schreiben. Tabelle 8 zeigt die entsprechende onDaF–DIALANG-Klassifikationstabelle.

Wie zu erkennen, wurden 39 Pbn im onDaF nach B1 eingestuft, im DIALANG-Subtest Schreiben aber nach A2. Übereinstimmende Einstufungen nach A2 liegen nur für 18 Pbn vor. Dies belegt die Vermutung, dass sich die Cut-Scores für die onDaF-Stufen A2 und B1 bei künftigen Untersuchungen mit einer größeren Datenbasis als revisionsbedürftig erweisen könnten.

Tabelle 8: onDaF–DIALANG-Klassifikationstabelle für den Subtest Schreiben (auf der Basis der ursprünglichen Cut-Scores für den onDaF)

	DIALANG-Stufe					
onDaF-Stufe	**A1**	**A2**	**B1**	**B2**	**C1/C2**	**Sum.**
unter A2	5	4	1	0	0	10
A2	7	18	6	0	0	31
B1	1	39	59	13	1	113
B2	0	0	28	23	4	55
C1 (od. höher)	0	0	1	7	2	10
Summe	13	61	95	43	7	219

Anmerkung: Anders als DIALANG differenziert onDaF nicht unterhalb von A2 bzw. oberhalb von C1. Aus Vergleichsgründen wurden die DIALANG-Stufen C1 und C2 zusammengefasst.

Die prozentuale Übereinstimmung für diese Klassifikationstabelle beträgt 48.9% (ausgedrückt als Anteil an der maximal möglichen Übereinstimmung ergeben sich 57.5%). Das (linear) gewichtete Kappa als Maß der zufallskorrigierten Übereinstimmung (Cohen, 1968; vgl. Bortz & Döring, 2006; von Eye & Mun, 2005) beläuft sich auf .45 ($p < .05$).

Die Randsummen sind sehr ungleich verteilt, vor allem sind die niedrigste und die höchste Kompetenzstufe nur sehr schwach vertreten. In einem solchen Fall ist der Anteil der beobachteten Übereinstimmung an der maximal möglichen Übereinstimmung ein empfohlener Indikator (vgl. Uebersax, 2006). Mit Werten zwischen 46% und 68% liegt dieser Indikator auf einem durchaus akzeptablen Niveau (es sei daran erinnert, dass hier allgemeine Einstufungen fertigkeits- bzw. bereichsspezifischen Einstufungen gegenübergestellt werden). Dennoch gab vor allem die beobachtete Abweichung im Bereich A2/B1 den Anstoß zu einer Revision der ursprünglichen Cut-Scores.

In einem iterativen Prozess wurden die Cut-Scores verändert, um eine bessere Übereinstimmung zwischen den onDaF- und den DIALANG-Einstufungen zu erzielen. Hiervon waren insbesondere die (zunächst nur provisorische) untere Grenze zu A2 sowie die untere und obere Grenze von B1 betroffen. Erstere wurde erhöht, die Breite des B1-Intervalls verringert. Diese Maßnahmen bewirkten leicht höhere Übereinstimmungsraten mit Werten zwischen 48% und 70% (Anteil der beobachteten an der möglichen Übereinstimmung). Die so modifizierten Cut-Scores bildeten den Schlusspunkt des geschilderten Standard-Setting- bzw. Validierungsprozesses.

5. Zusammenfassung und Diskussion

Nach gut zweijähriger Entwicklungsarbeit am TestDaF-Institut wurde der onDaF im Oktober 2006 erstmals für die weltweite Nutzung freigegeben. Seither (Stand: Dezember 2007) haben mehr als 10.000 Teilnehmer im In- und Ausland den onDaF abgelegt. Ein großer Teil entfällt auf Testungen an universitären Sprachenzentren. Diese verwenden den onDaF, um für ausländische Studierende den passenden Deutschkurs auszuwählen. Daneben nutzen immer mehr DAAD-Lektoren den onDaF als Sprachtest im Rahmen der Prüfung von Bewerbungen um ein DAAD-Stipendium für ein Studium in Deutschland. Schließlich dient der onDaF auch der Vorbereitung von Deutschlernern auf den TestDaF.[16]

Die bislang erhaltenen Rückmeldungen seitens der verschiedenen onDaF-Anwender sind sehr ermutigend. Insbesondere erfüllt der onDaF seine Hauptfunktion: die rasche und genaue Ermittlung des allgemeinen Sprachstands von Deutschlernern. Er erweist sich danach in der Testpraxis als ein ökonomisches und nützliches Messinstrument.

Ziel der vorliegenden Arbeit war die vertiefende Darstellung der theoretischen Grundlagen des onDaF sowie der Methoden seiner Konstruktion und Validierung. Die theoretischen Grundlagen betrafen hauptsächlich die spezifischen Herausforderungen internetgestützten Testens, den Aufbau einer kalibrierten Itembank sowie Konzepte und Verfahrensweisen zur Festlegung von Cut-Scores.

Hinsichtlich des ersten Punkts lassen sich die Eigenschaften des onDaF wie folgt zusammenfassen: (a) der onDaF ist ein serverzentrierter Test, d.h., alle für die Anwendung des Tests wichtigen Funktionen werden auf dem Server ausgeführt, die Clients (Teilnehmerrechner) stellen lediglich die Serverdaten auf dem Bildschirm dar und ermöglichen die Eingabe von Antworten über Tastatur und Maus, (b) der onDaF ist in der typischen Anwendung ein Medium-Stakes-Test, (c) die Testkontrolle erfolgt beim onDaF gemäß dem lizenzierten Modus, d.h., der onDaF darf nur an autorisierten Testabnahmestellen, die eine genaue Überwachung des Testablaufs vorsehen, durchgeführt werden, (d) die Testsicherheit wird durch serverseitige Speicherung aller Testmaterialien, strikte Kontrolle der Teilnehmeridentität, passwortgeschützte Teilnehmerdaten und ein Online-Verifizierungsmodul zur Prüfung der Echtheit von onDaF-Zertifikaten gewährleistet.

---

[16] Eine Variante des onDaF, die aus sechs Texten besteht und die Stufen A2 bis B2 abdeckt, kommt unter dem Namen onScreen mit wahlweise deutschen oder englischen Lückentexten als Instrument des Sprachscreenings im Vorfeld eines internationalen Studierfähigkeitstests („Test für Ausländische Studierende“ bzw. „Test for Academic Studies“, TestAS) zur Anwendung (www.testas.de).

Übergreifendes Ziel der Testentwicklung ist der sukzessive Aufbau einer kalibrierten Itembank. In dieser Itembank befinden sich erprobte Lückentexte zusammen mit ihren kategorialen, quantitativen und logischen Attributen. Das wichtigste quantitative Attribut ist die Schwierigkeit der Texte. Die Textschwierigkeiten werden auf der Basis des Ratingskalen-Rasch-Modells geschätzt. Alle Textschwierigkeiten werden dabei auf einer gemeinsamen Logitskala abgebildet. Das hierfür verwendete Design der Datenerhebung folgt dem Plan nichtäquivalenter Gruppen mit Ankertest (Ankertestplan). Als Ankertest fungieren zwei Lückentexte, die allen separat erprobten Sammlungen von je 10 Texten gemeinsam sind (interner Ankertest). Zur Schätzung der Textparameter auf der gemeinsamen latenten Dimension kommt eine simultane Schätzmethode zur Anwendung, d.h., die Schwierigkeitsparameter aller Texte werden in einer einzigen, simultanen Rasch-Analyse ermittelt.

Die Festlegung von Cut-Scores erfolgt beim onDaF in einer Weise, die sowohl die Besonderheiten des Testformats als auch die Ziele der Testanwendung (Einstufung der allgemeinen Deutschkenntnisse analog zu den GER-Niveaus A2 bis C1) berücksichtigt. Nach der neu entwickelten Prototypgruppenmethode benennen Sprachlehrer bzw. Leiter von Sprachkursen diejenigen Lerner, die sie als typische Vertreter einer definierten Niveaustufe betrachten. Anhand der Verteilungen, die sich für die Fähigkeitsschätzungen (Logits) der Lernerprototypen in den einzelnen Leistungskategorien ergeben, werden die Logitwerte bestimmt, die am besten zwischen je zwei benachbarten Kategorien zu unterscheiden erlauben. Statistische Verfahren, die hierfür eingesetzt werden, sind einmal das Median-Verfahren, d.h. die Berechnung des Medians der Logitwerte im Überlappungsintervall zwischen zwei benachbarten Kategorien, und zum anderen die binäre logistische Regression.

Die Ergebnisse der Analysen von Daten aus 18 weltweiten Erprobungen mit insgesamt 3.651 Probanden führten zur Aufnahme von 135 psychometrisch geeigneten Texten in die Itembank. Innerhalb der Erprobungssets resultierten sehr hohe Werte sowohl für die Reliabilität (mindestens .94) als auch für den Index der Klassenseparation (mindestens 5.69). In der simultanen Rasch-Analyse, die alle Teilnehmer und Texte zusammenfasste, lag die Klassenseparation bei einem Wert von 7.61, die Reliabilität belief sich auf .97. Es sind nach diesen Ergebnissen deutlich mehr Klassen von Teilnehmern zuverlässig unterscheidbar, als der onDaF Kompetenzstufen erfassen soll. Schließlich zeigten die Werte der Mean-Square-Fitstatistiken, dass die skalierten Texte das Kriterium der Eindimensionalität erfüllten.

Die von den Sprachlehrern bzw. Kursleitern abgegebenen Einstufungen der Lernerprototypen auf den Stufen A2 bis C1 waren eng korreliert mit den Schät-

zungen der Teilnehmerfähigkeiten, wie sie aus der simultanen Rasch-Analyse der Erprobungsdaten resultierten. Anders ausgedrückt, Teilnehmer, die Sprachlehrer als typische Vertreter einer unteren Kompetenzstufe benannt hatten, wiesen in der Rasch-Analyse signifikant niedrigere Fähigkeitsschätzungen auf als Teilnehmer, die Sprachlehrer als typische Vertreter einer höheren Kompetenzstufe eingestuft hatten. Die Einstufungen der Lernerprototypen bildeten daher eine geeignete Basis zur Anwendung der Prototypgruppenmethode des Standard-Settings.

Um die Validität des onDaF im Hinblick auf Einstufungen der Teilnehmerleistungen analog zu A2 bis C1 der GER-Skala genauer zu untersuchen, wurde ein Vergleich mit dem Deutschtest des Online-Sprachtestsystems DIALANG (Alderson, 2005; Alderson & Huhta, 2005) durchgeführt. DIALANG erlaubt Einstufungen in direkter Relation zu den GER-Stufen A1 bis C2, und zwar in den folgenden Fertigkeiten: Hörverstehen, Schreiben, Leseverstehen, Grammatik und Wortschatz. Im Sinne einer externen Validierung (Council of Europe, 2003; Figueras et al., 2005) fungierte DIALANG als Kriteriumstest, der es ermöglichen sollte, eine indirekte Beziehung des onDaF zum GER herzustellen.

Eine Reihe von Korrelations-, Regressions- und Faktorenanalysen erbrachte enge Zusammenhänge des onDaF mit den DIALANG-Subtests, insbesondere mit Schreiben, Grammatik und Wortschatz, und zwar sowohl auf der Ebene der Summenscores als auch auf der Ebene der Einstufungen. Bemerkenswert hoch waren aber auch die Zusammenhänge mit Hörverstehen. Mit der Ausnahme des Subtests Leseverstehen, der innerhalb des DIALANG-Deutschtests eine problematische Stellung einzunehmen schien, bewegten sich die Zusammenhänge auf einem ähnlich hohen Niveau wie jene, die Eckes & Grotjahn (2006a) zwischen einem verwandten C-Test (Papierversion) und den rezeptiven und produktiven Subtests des TestDaF beobachtet hatten.

Die nach onDaF und DIALANG vorgenommenen Einstufungen der Teilnehmerleistungen auf der globalen GER-Skala zeigten insgesamt ein zufrieden stellend hohes Maß an Übereinstimmung. In einem iterativen Prozess wurden beim onDaF die untere A2-Grenze und die Breite des Intervalls um B1 leicht modifiziert, um eine noch etwas höhere Übereinstimmung zwischen den onDaF- und den DIALANG-Einstufungen im unteren Skalenbereich zu erzielen.

Beachtung verdient auch der Befund, dass der onDaF mit allen Subtests aus DIALANG signifikant höher korrelierte als der in DIALANG selber eingesetzte Einstufungstest VSPT, der als reiner Wortschatztest konzipiert ist. Die substanziellen Korrelationen des onDaF mit den DIALANG-Subtests (zwischen .51 und .67, ausgenommen Leseverstehen) sind daher als ein weiterer Beleg für die Va-

lidität des onDaF als Instrument zur Messung allgemeiner Sprachkompetenz zu werten.

Künftige Untersuchungen zum onDaF werden sich den folgenden Themenbereichen widmen: (a) Beziehung zwischen dem onDaF und den Subtests des TestDaF, (b) Beziehung zwischen dem onDaF und einer Papierversion mit Texten aus der kalibrierten onDaF-Itembank, (c) Beziehung zwischen verschiedenen parallelen Testformen des onDaF.

Die Validierung eines Testverfahrens endet nicht mit dem Abschluss einer einzelnen empirischen Studie, sondern verlangt fortgesetzte Analysen zur Sicherung seiner psychometrischen Qualität auf einer möglichst breiten Basis. Nicht anders verhält es sich beim onDaF. Die in dieser Arbeit berichteten Forschungsergebnisse sind ein viel versprechender Anfang.

Danksagung

Ein Sprachtest wie der onDaF ist selbstverständlich nicht das Werk eines Einzelnen. Direkt oder indirekt an der Testentwicklung beteiligt waren Angehörige aller Abteilungen des TestDaF-Instituts. Die notwendigen IT-Strukturen und -Komponenten wurden von Florian Kuhlmann entwickelt. Duy Khuong Nguyen hat diese systematisch weiter ausgebaut und optimiert. Jolanta Lueg hat das Web-Design für den onDaF gestaltet und die onDaF-Seiten immer wieder auf den neuesten Stand gebracht. Zur Gestaltung des Web-Designs und zur Klärung einer Reihe praktischer Fragen der Client-Server-Relation hat Frank Weiss-Motz beigetragen. Achim Althaus hat Anregungen und Denkanstöße für möglichst einfache Lösungen komplexer angewandter Probleme gegeben. Andreas Kembügler hat Konzepte fürs Marketing entworfen und umgesetzt. Für eine zügige Durchführung der umfangreichen Erprobungen haben die weltweit tätigen DAAD-Lektoren sowie die Prüfungsbeauftragten des TestDaF-Instituts im In- und Ausland gesorgt. Ute Verwimp und Julia Haas haben Lückentexte erstellt, die Erprobungen logistisch betreut und die Teilnehmerantworten in schier endlosen Sitzungen am Rechner eingegeben. Beide haben außerdem an der Studie zur externen Validierung des onDaF mitgewirkt. Schließlich gilt auch den zahlreichen Studierenden, die als Testpersonen die Entwicklung des onDaF erst möglich gemacht haben, mein aufrichtiger Dank.

Literaturverzeichnis

Alderson, J. Charles. (2000). Technology in testing: The present and the future. *System, 28,* 593–603.

Alderson, J. Charles. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. Charles & Banerjee, Jayanti. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*, 79–113.

Alderson, J. Charles & Huhta, Ari. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*, 301–320.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, David. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Andrich, David. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105–113.

Angoff, William H. (1971). Scales, norms, and equivalent scores. In Robert L. Thorndike (Hrsg.), *Educational measurement* (2. Aufl., S. 508–600). Washington, DC: American Council on Education.

Ariel, Adelaide, van der Linden, Wim, J. & Veldkamp, Bernard P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement, 43*, 85–96.

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

Backhaus, Klaus, Erichson, Bernd, Plinke, Wulff & Weiber, Rolf. (2006). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (11. Aufl.). Berlin: Springer.

Barsalou, Lawrence W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Hillsdale, NJ: Erlbaum.

Bartram, Dave. (2006a). The internationalization of testing and new models of test delivery on the Internet. *International Journal of Testing, 6*, 121–131.

Bartram, Dave. (2006b). Testing on the Internet: Issues, challenges and opportunities in the field of occupational assessment. In Dave Bartram & Ronald K. Hambleton (Hrsg.), *Computer-based testing and the Internet: Issues and advances* (S. 13–37). Chichester, UK: Wiley.

Baur, Rupprecht & Spettmann, Melanie. (2005). Kompetenzstufen testen – leicht gemacht: C-Tests für DaF in der Praxis. In Dagmar Schäffer & Marjori Adamopoulou (Hrsg.), *Sprachen – Kulturen – Identität: Schule und Fortbildung für Europäer von morgen* (S. 149–161). Pallini, Griechenland: Ellinogermaniki Agogi.

Bisping, Meikel. (2006). Zur Validität von Computer-C-Tests. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (S. 147–166). Frankfurt: Lang.

Bisping, Meikel & Raatz, Ulrich. (2002). Sind computerisierte und Papier&Bleistift-Versionen des C-Tests äquivalent? In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Bd. 4, S. 131–155). Bochum: AKS-Verlag.

Bond, Trevor G. & Fox, Christine M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2. Aufl.). Mahwah, NJ: Erlbaum.

Bortz, Jürgen & Döring, Nicola (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Bühner, Markus. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2. Aufl.). München: Pearson Studium.

Chapelle, Carol A. & Douglas, Dan. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Choi, Inn-Chull, Kim, Kyoung S. & Boo, Jaeyool. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295–320.

Cizek, Gregory J. (2006). Standard setting. In Steven M. Downing & Thomas M. Haladyna (Hrsg.), *Handbook of test development* (S. 225–258). Mahwah, NJ: Erlbaum.

Cizek, Gregory J. & Bunch, Michael B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Clauser, Brian E. & Mazor, Kathleen M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31–44.

Cohen, Jacob. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cohen, Jacob, Cohen, Patricia, West, Stephen G. & Aiken, Leona S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3. Aufl.). Mahwah, NJ: Erlbaum.

Cook, Linda L. & Eignor, Daniel R. (1971). IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37–45.

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 Policy Division.

Coyne, Iain & Bartram, Dave. (2006). Design and development of the ITC guidelines on computer-based and Internet-delivered testing. *International Journal of Testing, 6,* 133–142.

Drasgow, Fritz & Mattern, Krista. (2006). New tests and new items: Opportunities and issues. In Dave Bartram & Ronald K. Hambleton (Hrsg.), *Computer-based testing and the Internet: Issues and advances* (S. 59–75). Chichester, UK: Wiley.

Drasgow, Fritz, Luecht, Richard M. & Bennett, Randy E. (2006). Technology and testing. In Robert L. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 471–515). Westport, CT: American Council on Education/Praeger.

Eckes, Thomas. (1991). *Psychologie der Begriffe: Strukturen des Wissens und Prozesse der Kategorisierung*. Göttingen: Hogrefe.

Eckes, Thomas. (1996). Begriffsbildung. In Joachim Hoffmann & Walter Kintsch (Hrsg.), *Lernen* (*Enzyklopädie der Psychologie, Kognition*, Bd. 7, S. 273–319). Göttingen: Hogrefe.

Eckes, Thomas. (2005). Evaluation von Beurteilungen: Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell. *Zeitschrift für Psychologie, 213,* 77–96.

Eckes, Thomas. (2006a, September). *Item banking for the onDaF: The Online Placement Test of German as a Foreign Language*. Paper presented at the 8th Anniversary of Korea Institute of Curriculum & Evaluation (KICE), Seoul, Republic of Korea.

Eckes, Thomas. (2006b). Rasch-Modelle zur C-Test-Skalierung. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (S. 1–44). Frankfurt: Lang.

Eckes, Thomas. (2007). Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen. *Diagnostica, 53,* 68–82.

Eckes, Thomas & Grotjahn, Rüdiger. (2006a). A closer look at the construct validity of C-tests. *Language Testing, 23,* 290–325.

Eckes, Thomas & Grotjahn, Rüdiger. (2006b). C-Tests als Anker für TestDaF: Rasch-Analysen mit dem kontinuierlichen Ratingskalen-Modell. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (S. 167–193). Frankfurt: Lang.

Embretson, Susan E. & Reise, Steven P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Figueras, Neus, North, Brian, Takala, Sauli, Verhelst, Norman & van Avermaet, Piet. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22,* 261–279.

Fischer, Gerhard H. (2007). Rasch models. In Calyampudi R. Rao & Sandip Sinharay (Hrsg.), *Psychometrics* (*Handbook of statistics*, Bd. 26, S. 515–585). Amsterdam: Elsevier.

Fischer, Johann. (2000). Diagnose aus dem Netz: DIALANG – ein Sprachtest im Internet. In Armin Wolff & Harald Tanzer (Hrsg.), *Sprache – Kultur – Politik* (S. 413–433). Regensburg: Fachverband Deutsch als Fremdsprache.

Folk, Valerie G. & Smith, Robert L. (2002). Models for delivery of CBTs. In Craig N. Mills, Maria T. Potenza, John J. Fremer & William C. Ward (Hrsg.), *Computer-based testing: Building the foundation for future asessments* (S. 41–66). Mahwah, NJ: Erlbaum.

Fulcher, Glenn. (2003). Interface design in computer-based language testing. *Language Testing, 20*, 384–408.

Germann, Ulrich & Grotjahn, Rüdiger. (1994). Das Lösen von C-Tests auf dem Computer: Eine Pilotuntersuchung zu den Bearbeitungsprozessen. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Bd. 2, S. 279–304). Bochum: Brockmeyer.

Grotjahn, Rüdiger. (2000). Testtheorie: Grundzüge und Anwendungen in der Praxis. In Armin Wolff & Harald Tanzer (Hrsg.), *Sprache – Kultur – Politik* (S. 304–341). Regensburg: Fachverband Deutsch als Fremdsprache.

Grotjahn, Rüdiger. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Bd. 4, S. 211–225). Bochum: AKS-Verlag.

Grotjahn, Rüdiger, Klein-Braley, Christine & Raatz, Ulrich. (2002). C-Tests: An overview. In James A. Coleman, Rüdiger Grotjahn & Ulrich Raatz (Hrsg.), *University language testing and the C-Test* (S. 93–114). Bochum: AKS-Verlag.

Häcker, Hartmut, Leutner, Detlev & Amelang, Manfred. (Hrsg.). (1998). *Standards für pädagogisches und psychologisches Testen*. Göttingen: Hogrefe.

Hambleton, Ronald K. & Pitoniak, Mary J. (2006). Setting performance standards. In Robert L. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, Ronald K. & Plake, Barbara S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–56.

Hanson, Bradley A. & Béguin, Anton A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3–24.

Harsch, Claudia & Schröder, Konrad. (2007). Textrekonstruktion: C-Test. In Bärbel Beck & Eckhard Klieme (Hrsg.), *Sprachliche Kompetenzen: Konzepte und Messung* (S. 212–225). Weinheim: Beltz.

Haswell, Richard H. (1998). Rubrics, prototypes, and exemplars: Categorization theory and systems of writing placement. *Assessing Writing, 5*, 231–268.

Henning, Grant. (1987). *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle & Heinle.

Henning, Grant. (1989). Meanings and implications of the principle of local independence. *Language Testing, 6*, 95–108.

Holland, Paul W. & Dorans, Neil J. (2006). Linking and equating. In Robert L. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 187–220). Westport, CT: American Council on Education/Praeger.

Huhta, Ari, Luoma, Sari, Oscarson, Mats, Sajavaara, Kari, Takala, Sauli & Teasdale, Alex. (2002). DIALANG: A diagnostic language assessment system for adult learners. In J. Charles Alderson (Hrsg.), *Common European Framework of Reference for Languages: Case studies* (S. 130–145). Strasbourg: Council of Europe.

Huynh, Huynh. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*, 19–20.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*, 93–114.

International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing, 6*, 143–171.

Jaeger, Richard M. (1989). Certification of student competence. In Robert L. Linn (Hrsg.), *Educational measurement* (3. Aufl., S. 485–514). New York: Macmillan.

Jamieson, Joan. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics, 25*, 228–242.

Kaftandjieva, Felianka. (2004). Standard setting. In Council of Europe (Hrsg.), *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section B, S. 1–43). Strasbourg: Language Policy Division.

Karantonis, Ana & Sireci, Stephen G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*, 4–12.

Kim, Seock-Ho & Cohen, Allan S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131–143.

Klein-Braley, Christine. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*, 47–84.

Klinck, Dorothea. (2002). *Computergestützte Diagnostik: Beeinflusst das Medium der Testbearbeitung die Testcharakteristika, die Testfairness oder das Erleben der Testsituation?* Göttingen: Hogrefe.

Klinck, Dorothea. (2006). Computerisierte Methoden. In Franz Petermann & Michael Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 226–232). Göttingen: Hogrefe.

Kolen, Michael J. & Brennan, Robert L. (2004). *Test equating, scaling, and linking: Methods and practices* (2. Aufl.). New York: Springer.

Koller, Gerhard & Zahn, Rosemary. (1996). Computer based construction and evaluation of C-tests. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 401–418). Bochum: Brockmeyer.

Kubinger, Klaus D. (1993). Testtheoretische Probleme der Computerdiagnostik. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 130–137.

Kubinger, Klaus D. (1999). Testtheorie: Probabilistische Modelle. In Reinhold S. Jäger & Franz Petermann (Hrsg.), *Psychologische Diagnostik: Ein Lehrbuch* (4. Aufl., S. 322–334). Weinheim: Psychologie Verlags Union.

Kubinger, Klaus D. (2006). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens.* Göttingen: Hogrefe.

Lee, Won-Chan. (2006, September). *Developing, maintaining, and calibrating item banks*. Paper presented at the 8th Anniversary of Korea Institute of Curriculum & Evaluation (KICE), Seoul, Republic of Korea.

Lee, Won-Chan, Song, Mi-Young & Kim, Jong-Pil. (2004, April). *An investigation of procedures for obtaining a common IRT scale*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Lehmann, Rainer H. (2003). Aspects of national and international surveys of student achievement in English as a foreign language. In Jutta Rymarczyk & Helga Haudeck (Hrsg.), *In search of the active learner: Untersuchungen zu Fremdsprachenunterricht, bilingualen und interdisziplinären Kontexten* (S. 155–162). Frankfurt: Lang.

Linacre, John M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, John M. (2007). *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs*. Chicago: Winsteps.com.

Linn, Robert L. (2006). The standards for educational and psychological testing: Guidance in test development. In Steven M. Downing & Thomas M. Haladyna (Hrsg.), *Handbook of test development* (S. 27–38). Mahwah, NJ: Erlbaum.

Little, David. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*, 167–190.

Livingston, Samuel A. & Zieky, Michael J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Livingston, Samuel A. & Zieky, Michael J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*, 121–141.

Lord, Frederic M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Luecht, Richard M. (2006). Operational issues in computer-based testing. In Dave Bartram & Ronald K. Hambleton (Hrsg.), *Computer-based testing and the Internet: Issues and advances* (S. 91–114). Chichester, UK: Wiley.

Mackiewicz, Wolfgang. (2001). *Wie man seine Sprachkenntnisse im Internet testet: Das EU-Projekt DIALANG*. Verfügbar unter: http://www.elfenbeinturm.net/archiv/2001/lern4.html

Masters, Geofferey N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Mitzel, Howard C., Lewis, Daniel M., Patz, Richard J. & Green, Donald R. (2001). The bookmark procedure: Psychological perspectives. In Gregory J. Cizek (Hrsg.), *Setting performance standards: Concepts, methods, and perspectives* (S. 249–281). Mahwah, NJ: Erlbaum.

Müller, Hans. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen: Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items*. Bern: Huber.

Murphy, Gregory L. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Myers, Jerome L. & Well, Arnold D. (2003). *Research design and statistical analysis* (2. Aufl.). Mahwah, NJ: Erlbaum.

Myford, Carol M. & Wolfe, Edward W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

North, Brian. (2004). Relating assessments, examinations, and courses to the CEF. In Keith Morrow (Hrsg.), *Insights from the Common European Framework* (S. 77–90). Oxford: Oxford University Press.

Ostini, Remo & Nering, Michael L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Rasch, Georg. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original erschienen 1960)

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

Röver, Carsten. (2001a). Web-based language testing. *Language Learning and Technology, 5*, 84–94.

Röver, Carsten. (2001b). Web-basiertes Testen fremdsprachlicher Fähigkeiten und Fertigkeiten. *Fremdsprachen Lehren und Lernen, 30*, 181–192.

Röver, Carsten. (2002). Web-based C-tests. In Rüdiger Grotjahn (Hrsg.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Bd. 4, S. 123–130). Bochum: AKS-Verlag.

Rost, Jürgen. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2. Aufl.). Bern: Huber.

Rudolf, Matthias & Müller, Johannes. (2004). *Multivariate Verfahren: Eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS*. Göttingen: Hogrefe.

Schermelleh-Engel, Karin, Kelava, Augustin & Moosbrugger, Helfried. (2006). Gütekriterien. In Franz Petermann & Michael Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 420–433). Göttingen: Hogrefe.

Schmeiser, Cynthia B. & Welch, Catherine J. (2006). Test development. In Robert L. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 307–353). Westport, CT: American Council on Education/Praeger.

Sigott, Günther. (2004). *Towards identifying the C-Test construct*. Frankfurt: Lang.

Singleton, David. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Sinharay, Sandip & Holland, Paul. (2006a). *Choice of anchor test in equating* (ETS Research Report, RR-06-35). Princeton, NJ: Educational Testing Service.

Sinharay, Sandip & Holland, Paul. (2006b). *The correlation between the scores of a test and an anchor test* (ETS Research Report, RR-06-04). Princeton, NJ: Educational Testing Service.

Sireci, Stephen G. (2001). Standard setting using cluster analysis. In Gregory J. Cizek (Hrsg.), *Setting performance standards: Concepts, methods, and perspectives* (S. 339–354). Mahwah, NJ: Erlbaum.

Sireci, Stephen G., Robin, Frédéric, Patelis, Thanos. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education, 12,* 301–325.

Smith, Richard M. (2004). Fit analysis in latent trait measurement models. In Everett V. Smith & Richard M. Smith (Hrsg.), *Introduction to Rasch measurement* (S. 73–92). Maple Grove, MN: JAM Press.

Stevens, James P. (2002). *Applied multivariate statistics for the social sciences* (4. Aufl.). Mahwah, NJ: Erlbaum.

Szabó, Gábor. (2008). *Applying item response theory in language test item bank building*. Frankfurt: Lang.

Uebersax, John. (2006). *Statistical methods for rater agreement*. Verfügbar unter: http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

Umar, Jahja. (1999). Item banking. In Geofferey N. Masters & John P. Keeves (Hrsg.), *Advances in measurement in educational research and assessment* (S. 207–219). Amsterdam: Pergamon.

Vale, C. David. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333–344.

Vale, C. David. (2006). Computerized item banking. In Steven M. Downing & Thomas M. Haladyna (Hrsg.), *Handbook of test development* (S. 261–285). Mahwah, NJ: Erlbaum.

van der Linden, Wim J. (2005). *Linear models for optimal test design*. New York: Springer.

von Davier, Alina A., Holland, Paul W. & Thayer, Dorothy T. (2004). *The kernel method of test equating*. New York: Springer.

von der Handt, Gerhard. (2001). DIALANG – ein diagnostisches Online-Testverfahren (Schwerpunkt Hörverstehen). *Fremdsprachen Lehren und Lernen, 30,* 167–180.

von Eye, Alexander & Mun, Eun Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah, NJ: Erlbaum.

Wainer, Howard, Bradlow, Eric T. & Wang, Xiaohui. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, Howard & Kiely, Gerard L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wilson, Mark. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12,* 353–364.

Wilson, Mark & Adams, Raymond J. (1995). Rasch models for item bundles. *Psychometrika, 60,* 181–198.

Wingersky, Marilyn S. & Lord, Frederic M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wolfe, Edward W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1*, 409–434.

Wright, Benjamin D. & Masters, Geofferey N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, Benjamin D. & Masters, Geofferey N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16*(3), 888.

Wright, Benjamin D. & Stone, Mark H. (1999). *Measurement essentials* (2. Aufl.). Wilmington, DE: Wide Range.

Yen, Wendy M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–214.

Yen, Wendy M. & Fitzpatrick, Anne R. (2006). Item response theory. In Robert L. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 111–153). Westport, CT: American Council on Education/Praeger.

Zieky, Michael J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In Gregory J. Cizek (Hrsg.), *Setting performance standards: Concepts, methods, and perspectives* (S. 19–51). Mahwah, NJ: Erlbaum.

Zieky, Michael & Perie, Marianne. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.

Zydatiß, Wolfgang. (2005). *Bildungsstandards und Kompetenzniveaus im Englischunterricht: Konzepte, Empirie, Kritik und Konsequenzen*. Frankfurt: Lang.

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn and Günther Sigott

Vol. 1 Günther Sigott: Towards Identifying the C-Test Construct. 2004.

Vol. 2 Carsten Röver. Testing ESL Pragmatics. Development and Validation of a Web-Based Assessment Battery. 2005.

Vol. 3 Tom Lumley: Assessing Second Language Writing. The Rater's Perspective. 2005.

Vol. 4 Annie Brown: Interviewer Variability in Oral Proficiency Interviews. 2005.

Vol. 5 Jianda Liu: Measuring Interlanguage Pragmatic Knowledge of EFL Learners. 2006.

Vol. 6 Rüdiger Grotjahn (Hrsg./ed.): Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, Empirical Research, Applications. 2006.

Vol. 7 Vivien Berry: Personality Differences and Oral Test Performance. 2007.

Vol. 8 John O'Dwyer: Formative Evaluation for Organisational Learning. A Case Study of the Management of a Process of Curriculum Development. 2008.

Vol. 9 Aek Phakiti: Strategic Competence and EFL Reading Test Performance. A Structural Equation Modeling Approach. 2007.

Vol. 10 Gábor Szabó: Applying Item Response Theory in Language Test Item Bank Building. 2008.

Vol. 11 John M. Norris: Validity Evaluation in Language Assessment. 2008.

Vol. 12 Barry O'Sullivan: Modelling Performance in Tests of Spoken Language. 2008.

Vol. 13 Annie Brown / Kathryn Hill (eds.): Tasks and Criteria in Performance Assessment. Proceedings of the 28th Language Testing Research Colloquium. 2009.

Vol. 14 Ildikó Csépes: Measuring Oral Proficiency through Paired-Task Performance. 2009.

Vol. 15 Dina Tsagari: The Complexity of Test Washback. An Empirical Study. 2009.

Vol. 16 Spiros Papageorgiou: Setting Performance Standards in Europe. The Judges' Contribution to Relating Language Examinations to the Common European Framework of Reference. 2009.

Vol. 17 Ute Knoch: Diagnostic Writing Assessment. The Development and Validation of a Rating Scale. 2009.

Vol. 18 Rüdiger Grotjahn (Hrsg./ed.): Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research. 2010.

www.peterlang.de

C-Tests bestehen aus mehreren kurzen Texten, in denen fehlende Wortteile zu rekonstruieren sind. C-Tests haben hervorragende psychometrische Eigenschaften und werden in einer Vielzahl von Kontexten zur validen und ökonomischen Messung allgemeiner Sprachkompetenz eingesetzt. Dieser Sammelband illustriert den aktuellen Stand der C-Test-Forschung – mit einem Schwerpunkt auf folgenden Aspekten: Validität von C-Tests; Rasch-Modelle für C-Test-Daten; Zuordnung von C-Test-Ergebnissen zum Gemeinsamen europäischen Referenzrahmen für Sprachen.

C-Tests consist of several short texts in which the missing parts of words have to be reconstructed. C-Tests have excellent psychometric properties and are used in many contexts as valid and economical tests of general language proficiency. This collection of papers illustrates the state of the art of C-Test research, with a special focus on the following issues: validity of C-Tests; Rasch measurement models for C-Test data; relating C-Test results to the Common European Framework of Reference for Languages.

Rüdiger Grotjahn ist Professor am Seminar für Sprachlehrforschung der Universität Bochum. Seine Hauptforschungsgebiete sind Sprachtests, Forschungsmethodologie und individuelle Unterschiede beim Fremdsprachenlernen. Ein Schwerpunkt seiner umfangreichen Publikationstätigkeit liegt im Bereich des C-Tests.

Rüdiger Grotjahn is a professor at the Department of Second Language Research, University of Bochum. His main research interests are in language testing, research methodology, and the study of individual differences in language learning. His numerous publications include extensive work on the C-Test.

www.peterlang.de