Rüdiger Grotjahn, Bochum

Testtheorie: Grundzüge und Anwendungen in der Praxis

1. Einleitung

Im Zentrum meines Beitrags steht die Test**theorie**. Dies führt zu einem gewissen Dilemma: Die Testtheorie und mit einer beträchtlichen Verzögerung auch die **Sprach**testtheorie hat sich zu einer für nur noch wenige Spezialisten verständlichen Wissenschaft mit einem hohen Mathematisierungsgrad entwickelt. Dies belegt eine Vielzahl von insbesondere jüngeren Beiträgen in der Zeitschrift "Language Testing" sowie auch eine Reihe im Literaturverzeichnis genannter Publikationen.

Der vorliegende Beitrag wendet sich nicht an die wenigen Spezialisten, sondern vor allem an Praktiker. Praktiker fragen jedoch häufig „Wofür brauchen wir überhaupt Testtheorie?". Wird dann noch deutlich, daß Testtheorie sehr viel mit Messen, Statistik und Mathematik zu tun hat, stößt man nicht selten auf massive Ablehnung und Verweigerung.

Zustimmung erhält man allerdings, wenn man fordert, daß ein Test gültige und faire Entscheidungen hinsichtlich der getesteten Lerner ermöglichen soll. Ohne ein gewisses Maß an Testtheorie ist jedoch weder die Konstruktion eines gültigen und fairen Tests noch eine begründete Bewertung vorliegender Testverfahren möglich. Auch ein Praktiker benötigt deshalb Kenntnisse in der Testtheorie.

2. Die Begriffe „Test", „Testen", „Item" und „Aufgabe"

Die für mein Thema zentralen Begriffe „Test", „Testen", „Item" und „Aufgabe" werden in unterschiedlicher Weise verwendet. Es sollen deshalb zunächst einige Begriffsklärungen vorgenommen werden.

Ich werde im folgenden von einem weiten Verständnis des Begriffs „**Test**" ausgehen. Unter „Test" soll jegliches Prüfungsverfahren gefaßt werden, das Individuen unter kontrollierten Bedingungen zu bestimmten Handlungs- und Verhaltensweisen veranlaßt, die Rückschlüsse ermöglichen sollen auf zugrundeliegende Persönlichkeitsmerkmale wie Sprachfähigkeit oder Wissensstrukturen, auf spezifische Fertigkeiten wie das Schreiben von fremdsprachigen Zusammenfassungen und/oder auf den Stand in bezug auf einen bestimmten Maßstab, wie z.B. Lehrziele oder Leistung einer Vergleichsgruppe. Diese – zugegebenermaßen komplexe – Definition schließt sowohl standardisierte Prüfungen, wie z.B. des Goethe-Instituts, als auch die sog. informellen Tests ein, die häufig von Lehrern ad hoc für einen ganz spezifischen Zweck konzipiert werden. Sie ist damit deutlich weiter als die bekannte Definition von Lienert/Raatz (1994: 1), in der der Begriff „Test" auf „**wissenschaftliche Routineverfahren**" eingeschränkt wird.

Ein typisches Merkmal von Tests ist, daß sie kürzere und/oder einfachere Ersatzverfahren für ein vollständigeres und aufwendigeres Verfahren darstellen. So ist z.B. ein zweistündiger, als Gruppenprüfung durchgeführter fremdsprachlicher Studieneingangstest ein ökonomischer, d.h. kosten- und arbeitssparender Ersatz, für eine langwierige individuelle Diagnose des Sprachstandes jedes einzelnen Probanden im Studienverlauf. Nicht unter das explizierte Verständnis von „Test" fällt z.B. die Beurteilung der sprachlichen Leistung anhand von sog. Portfolios, d.h. Lerndossiers. Im englischen Sprachraum wird hier im übrigen auch nicht von "testing", sondern von "assessment" oder auch "evaluation" gesprochen.

Unter den Terminus „**Item**" fällt jedes Einzelelement eines Tests, das eine bestimmte Reaktion auf seiten der Kandidaten hervorrufen soll und das getrennt von den übrigen Elementen bewertet wird. Beispiele sind: einzelne Lücken in einem Cloze Test oder einzelne Multiple-Choice-Aufgaben (Mehrfachwahlaufgaben). Daneben wird der Begriff auch für eine zusammengehörende Gruppe von Items verwendet, so z.B. für einen einzelnen Text eines C-Tests. Der Punktwert für das Item ergibt sich dann aus der Summe der Punktwerte der zu dem jeweiligen Text gehörenden Einzelitems (d.h. den Lücken).

In seiner **engen** Bedeutung steht „Aufgabe" für „Item". Der Terminus „Mehrfachwahlaufgabe" ist ein bekanntes Beispiel für die enge Variante. In der im folgenden zugrundegelegten **weiten** Bedeutung steht „Aufgabe" für „ein dem Probanden zur Lösung vorgelegtes Problem". Damit kann eine Aufgabe dem gesamten vom Lerner zu bearbeitenden Test, einzelnen Untertests, Aufgabentypen oder auch spezifischen Items entsprechen.

3. Klassische vs. probabilistische Testtheorie

Sowohl im Hinblick auf die Entwicklung als auch bezüglich der Analyse von Tests ist der Unterschied zwischen der sog. klassischen Testtheorie und der sog. probabilistischen Testtheorie von zentraler Bedeutung. Beide Theorien

Wolff, Armin/Tänzer, Harald (eds.): *Sprache – Kultur – Politik. Beiträge der 27. Jahrestagung Deutsch als Fremdsprache vom 3.-5- Juni 1999 an der Universität Regensburg.* Universität Regensburg: Fachverband Deutsch als Fremdsprache (= Materialien Deutsch als Fremdsprache, Bd. 53) 2000

gehen von der Annahme aus, daß Messungen und damit auch Tests grundsätzlich **fehlerbehaftet** sind; sie unterscheiden sich jedoch grundlegend in der Art, wie sie der Fehlerbehaftetheit von Messungen Rechnung tragen.

In der klassischen Testtheorie wird der beobachtete Punktwert eines Probanden auf zwei additive Komponenten zurückgeführt: (1) den nicht direkt beobachtbaren **wahren** Wert des Probanden und (2) einen vom wahren Wert unabhängigen **Meßfehler**. Sodann werden hinsichtlich des Meßfehlers eine Reihe von Annahmen gemacht und zwar insbesondere, daß die Meßfehler sich auf die Dauer gegenseitig ausgleichen, d.h. daß ihr Mittel- bzw. Erwartungswert 0 ist, und daß sie zudem vom wahren Wert unabhängig sind. Ist der Erwartungswert der Meßfehler ungleich 0, dann liegt ein **systematischer Fehler** vor (vgl. z.B. Baker 1997; Krauth 1995; Lienert/Raatz 1994; Stumpf 1996). Dies ist z.B. der Fall, wenn die sprachliche Leistung eines Probanden in einem mündlichen Interview aufgrund von äußeren Faktoren, wie z.B. dem netten und freundlichen Verhalten des Probanden, immer wieder systematisch überschätzt wird.

Die klassische Testtheorie ist in erster Linie eine Theorie über Eigenschaften des Meßfehlers. Die probabilistische Testtheorie trägt dagegen der Meßfehlerproblematik dadurch Rechnung, daß sie davon ausgeht, daß die Antworten der Probanden probabilistischen Charakter haben, d.h. Wahrscheinlichkeitsgesetzen folgen. So wird z.B im bekannten Rasch-Modell angenommen, daß eine Beziehung besteht zwischen der Wahrscheinlichkeit, daß eine Person eine Aufgabe löst und der Differenz zwischen der Fähigkeit der Person und der Schwierigkeit der Aufgabe. Je größer die Fähigkeit der Person in Relation zur Schwierigkeit der Aufgabe ist, desto größer wird die Wahrscheinlichkeit, daß die Person die Aufgabe korrekt löst. Entspricht die Fähigkeit der Person genau der Schwierigkeit der Aufgabe, ist die Wahrscheinlichkeit für eine korrekte Lösung 50%. Übersteigt die Fähigkeit der Person die Schwierigkeit der Aufgabe, ist die Wahrscheinlichkeit größer als 50%. Im umgekehrten Fall beträgt die Wahrscheinlichkeit weniger als 50%. Die zu messende Eigenschaft wird häufig als "latent trait", d.h. als nicht beobachtbare, zugrundeliegende Fähigkeit bzw. Verhaltensdisposition aufgefaßt. Die nicht direkt beobachtbare fremdsprachliche Kompetenz eines Lerners ist ein solcher "latent trait". Den "latent trait" gilt es anhand des manifesten Verhaltens, d.h. anhand der beobachteten Reaktionen auf die Testaufgaben, zu messen. Entsprechend wird dieser Ansatz auch unter "Latent-Trait-Modelle" abgehandelt. Insbesondere in der englischsprachigen Literatur wird auch die Bezeichnung "Item Response Theory" (IRT) verwendet (vgl. z.B. Fischer/Molennaar 1995; Rost 1996; Baker 1997; Pollitt 1997; van der Linden/Hambleton 1997).

Verglichen mit der klassischen Testtheorie hat die probabilistische Testtheorie eine Reihe von Vorteilen. Ich nenne einige wichtige:

1. Die auf der Basis der probabilistischen Testtheorie ermittelten Kennwerte sind von der jeweiligen Stichprobe bzw. Referenzpopulation **un**abhängig. In der klassischen Testtheorie sind die berechneten statistischen Kennwerte, wie z.B. Schwierigkeitsindizes oder Reliabilitätskoeffizienten, dagegen von der Ausprägung der zu messenden Fähigkeit in der Population bzw. Stichprobe abhängig. So kann z.B. ein und dasselbe Item ganz unterschiedliche Schwierigkeitswerte erhalten, je nachdem wie die zu messende Fähigkeit in der jeweiligen Stichprobe ausgeprägt ist. Diese Populations**abhängigkeit** ist ein entscheidender Nachteil der klassischen Testtheorie. Populations**unabhängigkeit** der Kennwerte der probabilistischen Testtheorie impliziert allerdings nicht „die Übertragbarkeit von Ergebnissen von einer Population auf eine beliebige andere" (Fischer 1974: 134). So kann z.B. aufgrund von kulturellen Differenzen oder auch aufgrund der jeweiligen Muttersprache die Gültigkeit eines Fremdsprachentests in bezug auf bestimmte Probandengruppen stark eingeschränkt sein. Dieser Tatsache gilt es bei der Entwicklung eines multinational eingesetzten Tests, wie des neuen „Test Deutsch als Fremdsprache" (TESTDAF), Rechnung zu tragen.[1]
2. Die probabilistische Testtheorie ermöglicht begründete Aussagen zur Dimensionalität und zum Skalenniveau des zu messenden Merkmals.
3. Verglichen mit der klassischen Testtheorie ermöglicht die probabilistische Testtheorie fundiertere und differenziertere Aussagen zu den Fähigkeiten einzelner Individuen.
4. Die probabilistische Testtheorie ist zur Erstellung von Itembanken und zur Konstruktion von adaptiven Tests weit geeigneter als die klassische Testtheorie.

---

[1] TESTDAF richtet sich an ausländische Studienbewerber, die ein Studium in Deutschland beginnen wollen. TESTDAF ist in seiner Konzeption vergleichbar mit dem IELTS (*International English Language Testing System*) und dem TOEFL (*Test of English as a Foreign Language*). TESTDAF wird seit dem 1.8.1998 im Auftrag des DAAD von einem Konsortium entwikkelt, das aus folgenden Institutionen besteht: FernUniversität Hagen, Goethe-Institut, Carl Duisberg Centren, Seminar für Sprachlehrforschung der Ruhr-Universität Bochum. Der Verfasser ist Mitglied des Entwicklungskonsortiums.

Die Mehrzahl der in der Praxis eingesetzten Tests beruht allerdings immer noch auf den Prinzipien der klassischen Testtheorie. Dies hat eine Reihe von eher praktischen Gründen:

1. Die Anwendung der probabilistischen Testtheorie setzt im Vergleich zur klassischen Testtheorie relativ große Stichproben voraus. Diese sind jedoch in der Praxis häufig nicht gegeben.
2. Eine Analyse von Testdaten mit Hilfe der probabilistische Testtheorie ist ohne spezielle Software nicht möglich. Die entsprechende Software steht zumeist nicht unmittelbar zur Verfügung. Dies gilt zumindest für die Rechenzentren deutscher Universitäten.
3. Die probabilistische Testtheorie ist mathematisch sehr komplex. Auch viele Sprachtester sind deshalb nicht hinreichend mit ihr vertraut.

Im Rahmen großer internationaler Sprachtests und bei professionellen Testentwicklungsorganisationen findet jedoch die probabilistische Testtheorie immer häufiger Verwendung. Dies gilt zumindest in bezug auf Nordamerika, England, Australien oder auch die Niederlande, wo der Einsatz von probabilistischen Testmodellen schon eine relativ lange Tradition hat. In der deutschen Sprachtestforschung ist die probabilistische Testtheorie dagegen bisher kaum eingesetzt worden. Eine aktuelle Ausnahme ist die Verwendung des einparametrischen Rasch-Modells im Rahmen der Entwicklung des TESTDAF.

4. Gütekriterien

Tests müssen bestimmten Gütekriterien genügen. In der klassischen Testtheorie werden zumeist folgende Hauptgütekriterien genannt: Objektivität, Reliabilität und Validität. Daneben finden sich als weitere Kriterien z.B. die Ökonomie, die Praktikabilität, die Nützlichkeit, die Fairness, die Transparenz[1] oder auch die Normierung von Tests (vgl. z.B. Ingenkamp 1985: 34-43; Trim/North 1992; Lienert/Raatz 1994: 7-14; Bachman/Palmer 1996; Kieweg 1999). Während für Lienert/Raatz (1994) das Merkmal der Nützlichkeit ledig-

[1] Das Kriterium der Transparenz spielt u.a. in den Ausführungen des Europarats zum Fremdsprachenunterricht und zur Zertifizierung eine wichtige Rolle (vgl. z.B. Trim/North 1992). Auf die Notwendigkeit, Lernzielkontrollen für die betroffenen Schüler hinreichend transparent zu gestalten, hat jüngst Kieweg (1999: 10) hingewiesen:

„Die in einer Prüfungssituation latent vorhandenen Angstwerte können immer dann etwas reduziert werden, wenn der Schüler eine Lernzielkontrolle erledigen muss, die das Kriterium der ausreichenden **Transparenz** erfüllt. Eine Lernzielkontrolle ist für Schüler dann transparent,
- wenn die verschiedenen Kontrollverfahren bekannt sind,
- wenn die darin enthaltenen Aufgaben und Fragen unmisserständlich klar formuliert sind,
- wenn die zu erreichenden Punkte deutlich ausgewiesen sind und
- wenn der Beurteilungsschlüssel angegeben ist." (Hervorhebung R. G.)

lich ein Nebengütekriterium ist, sehen andere Autoren die Nützlichkeit als sehr wichtiges oder sogar als letztlich entscheidendes Kriterium für die Qualität eines Tests an. Schließlich wird vor allem auch im Sprachtestbereich die Authentizität der Testaufgaben als weiteres wichtiges Kriterium angeführt.

Ich gehe zunächst auf die Kriterien „Objektivität", „Reliabilität" und „Validität" ein und beschränke mich dabei auf eine Charakterisierung aus der Sicht der **klassischen** Testtheorie.

Das Gütekriterium der **Objektivität** bezieht sich nach Lienert/Raatz (1994: 7) auf den „Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind." Eine zentrale Voraussetzung für eine zufriedenstellende Objektivität ist die **Standardisierung** (vereinheitlichende Festlegung) des Vorgehens bei der Durchführung, Auswertung und Interpretation eines Tests. Kommen – z.B. infolge einer Standardisierung – verschiedene Untersucher bei den gleichen Kandidaten zu den gleichen Ergebnissen, dann ist ein Test vollständig objektiv. Umgekehrt gilt nach Ingenkamp (1985: 34): „Wenn wir bei einem Meßergebnis nicht mehr unterscheiden können, wie weit es Merkmale des Gemessenen oder des Messenden kennzeichnet, wenn wir annehmen müssen, daß ein anderer Beobachter zu einem ganz anderen Ergebnis gekommen wäre, dann können wir aus diesem Meßergebnis keine Aussagen und Folgerungen ableiten, die von über den Zufall hinausgehender Bedeutung sind." Entsprechend kommt Ingenkamp (1985: 36) auch zu folgender Feststellung: „**Wer darauf verzichtet, sich um Objektivität zu bemühen, der überläßt letzten Endes unkontrollierter Willkür das Feld**" (Hervorhebung im Original).

Es wird häufig zwischen Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität unterschieden (vgl. Ingenkamp 1985: 34ff.; Lienert/Raatz 1994: 8).

Die **Durchführungsobjektivität** bezieht sich auf die Unabhängigkeit der Testergebnisse vom Verhalten des Untersuchers während der Testdurchführung. Der Grad der Durchführungsobjektivität hängt in erster Linie von der Güte der Standardisierung der Testdurchführung ab.

Die **Auswertungsobjektivität** betrifft die Auswertung der registrierten Reaktionen der Kandidaten nach vorgegebenen Regeln. So ist z.B. ein Lückentest, bei dem lediglich solche Lösungen als richtig gelten, die in einer vorliegenden Liste aufgeführt sind, vollständig auswertungsobjektiv.

Die **Interpretationsobjektivität** bezieht sich auf den Grad der Unabhängigkeit der Interpretation der ausgewerteten Testergebnisse von der Person des interpretierenden Testbenutzers. Vollkommene Interpretationsobjektivität liegt vor, wenn gleiche Auswertungsergebnisse gleich interpretiert werden, d.h. wenn aus ihnen die gleichen Schlüsse gezogen werden. Dies ist z.B. gewährleistet, wenn die Auswertung einen numerischen Wert liefert, der die Position des Kandidaten auf einer Skala relativ zu den übrigen Kandidaten festlegt. In einem solchen Fall ist die Auswertung nicht nur eine notwendige, sondern zugleich auch eine hinreichende Voraussetzung für eine Aussage über den Kandidaten. So hat z.B. ein standardisierter Papier-Bleistift-Test im Multiple-Choice-Format, der die Kandidaten lediglich in eine Rangordnung bringen soll, sowohl eine hohe Interpretationsobjektivität als auch eine hohe Durchführungs- und Auswertungsobjektivität. Je unterschiedlicher und je zahlreicher jedoch die von einem Test gelieferten Informationen sind, desto schwieriger ist es in der Regel, die Testergebnisse objektiv zu interpretieren. Dies ist z.B. der Fall, wenn die Ergebnisse nicht als einzelner Punktwert, sondern in Form eines komplexen mehrdimensionalen Profils vorliegen.[1]

Ein aktuelles Beispiel für einen unzureichend standardisierten und damit auch nicht hinreichend objektiven Test ist die „Deutsche Sprachprüfung für den Hochschulzugang" (DSH). Die DSH ist weder hinreichend durchführungsobjektiv, noch hinreichend auswertungsobjektiv, noch hinreichend interpretationsobjektiv, da die Durchführungs- Auswertungs- und Interpretationsmodalitäten sowohl innerhalb einer bestimmten Institution als auch von Institution zu Institution erheblich differieren können.

Die **Reliabilität** – auch als **Zuverlässigkeit** bezeichnet – bezieht sich auf die Genauigkeit, mit der Testergebnisse eine Eigenschaft erfassen, unabhängig davon, ob der Test wirklich die Eigenschaft mißt, die gemessen werden soll. So kann z.B. ein als Sprachtest konzipiertes Verfahren hoch reliabel sein, auch dann, wenn es in Wirklichkeit etwas gänzlich anderes mißt. In der Regel gilt zumindest prinzipiell im Fall einer eindimensionalen Eigenschaft: Je größer die Zahl der Items ist, die die Eigenschaft messen, desto zuverlässiger wird die Eigenschaft gemessen. Dies ist einer der Gründe, warum professionelle Tests aus relativ vielen Items bestehen.

[1] Dies bedeutet keineswegs, daß man darauf verzichten sollte, Testergebnisse in Form von differenzierten, mehrdimensionalen Profilen darzustellen. M.E. wird dies viel zu selten getan. Nichtsdestoweniger sollte man sich jedoch über mögliche Schwierigkeiten bei der Interpretation entsprechender Profile im klaren sein (vgl. hierzu auch Brindley 1998).

Lassen sich Testergebnisse nur sehr ungenau reproduzieren, d.h., bekommt man bei einer Testwiederholung deutlich abweichende Ergebnisse, ist dies ein Hinweis auf eine unbefriedigende Reliabilität des Verfahrens – allerdings nur unter der Voraussetzung, daß sich die zu messende Eigenschaft, z.B. durch zwischenzeitlichen Unterricht, nicht verändert hat. Dieser sich auf die zeitliche Stabilität der Testergebnisse beziehende Typ von Reliabilität wird als **Retestreliabilität** oder **Testwiederholungsreliabilität** bezeichnet.

Verfügt man über zwei parallele, d.h. äquivalente Formen ein und desselben Tests und setzt die Parallelformen in ein und derselben Stichprobe ein, dann kann man die Reliabilität über die Korrelation der beiden Paralleltests ermitteln. Man spricht deshalb auch von der **Paralleltestreliabilität**. Verglichen mit der Retestmethode ist die Paralleltestmethode das bessere Verfahren der Reliabilitätsbestimmung, da die Möglichkeit der zwischenzeitlichen Änderung des zu messenden Merkmals entfällt. Voraussetzung ist jedoch, daß man nachgewiesen hat, daß die beiden Testformen äquivalent sind, d.h. das gleiche Konstrukt messen.

In der Praxis verfügt man zumeist nicht über parallele Formen eines Tests. Auch der wiederholte Einsatz ein und desselben Test ist häufig nicht möglich. Man ist deshalb gezwungen, die Reliabilität auf der Basis eines einzigen Testdurchgangs zu ermitteln. Dazu teilt man den Test in zwei äquivalente Hälften oder falls möglich in so viele Teile wie es Items gibt, und schätzt die Reliabilität über die Korrelation der Teile untereinander. Ist die mittlere Korrelation hoch, dann ist der Gesamttest reliabel im Sinne einer **internen Konsistenz** der Messungen. Konkret bedeutet dies, daß die einzelnen Aufgaben die Testteilnehmer jeweils in eine ähnliche Rangfolge bringen. Eine Voraussetzung dieses Verfahrens ist allerdings wiederum die Äquivalenz der einzelnen Teile bzw. Items. Cronbachs Reliabilitätskoeffizient Alpha ist ein bekanntes Beispiel für einen Konsistenzkoeffizienten.

Ist die Reliabilität eines Tests gleich 0, dann mißt der Test völlig unzuverlässig. Beträgt die Reliabilität 1, mißt der Test absolut zuverlässig. In der Regel wird gefordert, daß ein Test, der zur **Differenzierung zwischen einzelnen Individuen** eingesetzt wird, eine Reliabilität von mindestens 0,9 aufweisen sollte. Soll jedoch lediglich ein **globaler Vergleich zwischen Gruppen** z.B. für Forschungszwecke durchgeführt werden, dann gilt zumeist eine Reliabilität von ca. 0,6 als ausreichend. Hier wird deutlich, daß die Reliabilität stets im Hinblick auf die jeweilige Anwendung eines Tests zu bewerten ist.

Wegen der Populationsabhängigkeit der Kennwerte der klassischen Testtheorie charakterisiert die Reliabilität allerdings lediglich die Zuverlässigkeit eines Tests in bezug auf eine bestimmte Population; eine populationsunabhängige Aussage über die Genauigkeit der Messung einer Person ist auf der Basis des Reliabilitätskonzepts der klassischen Testtheorie nicht möglich.

Das für die meisten Autoren entscheidende Gütekriterium eines Test ist dessen **Validität**. Die Validität – auch als **Gültigkeit** bezeichnet – bezieht sich auf das Ausmaß, in dem ein Test das erfaßt, was er erfassen soll. Hervorzuheben ist, daß das Gütekriterium der Validität stets in Abhängigkeit von der spezifischen Verwendung eines Tests zu sehen ist. Einer Aussage wie „Dieser Test ist valide" ist deshalb ohne Angabe der Verwendung des Tests mit größtem Mißtrauen zu begegnen.

Es wird u.a. zwischen folgenden Validitätstypen unterschieden:

1. Inhaltsvalidität (Kontentvalidität)
2. kriterienbezogene Validität
   a) Übereinstimmungsvalidität
   b) prädiktive Validität
3. Augenscheinvalidität (*face validity*)
4. Konstruktvalidität

Die **Inhaltsvalidität** gibt das Ausmaß an, in dem die Testaufgaben geeignet sind, z.B. bestimmte Aspekte eines Lernstoffs oder auch bestimmte Verhaltensweisen zu erfassen. Das Kriterium bezieht sich damit auf die Relevanz und Repräsentativität des Testinhalts in bezug auf den zu messenden Bereich. Die Inhaltsvalidität wird anhand von Expertenurteilen ermittelt und ist ein relativ problematisches Kriterium. Aus empirischen Untersuchungen weiß man, daß Experten erheblich in der Einschätzung der Inhaltsvalidität differieren können. Außerdem bleiben bei der Beurteilung der Inhaltsvalidität die tatsächlichen Reaktionen der Kandidaten auf die Testaufgaben unberücksichtigt. Manche Autoren sprechen deshalb auch von logischer Validität. Insgesamt gesehen ist die Inhaltsvalidität bestenfalls eine notwendige, nicht jedoch eine hinreichende Voraussetzung für die Validität eines Tests. Um eine zufriedenstellende Inhaltsvalidität (inhaltliche Repräsentativität) zu erreichen, benötigen wir bei einem heterogenen Merkmal, das – wie z.B. kommunikative Kompetenz – aus einer Vielzahl von Unterdimensionen besteht, weit mehr Items, als bei einem homogenen Merkmal. Auch aus diesem Grunde bestehen viele Tests aus einer großen Zahl von Einzelitems.

Bei der **kriterienbezogenen Validität** wird geprüft, inwieweit die Testergebnisse mit einem unabhängigen Außenkriterium, wie z.B. einem anderen Test, übereinstimmen. Der Grad der Übereinstimmung wird in der Regel anhand eines Korrelationskoeffizienten gemessen. Die Höhe der Korrelation charakterisiert u.a. das Ausmaß, in dem die Tests das gleiche Konstrukt messen. So bedeutet z.B. eine Korrelation von 0,5 zwischen dem Testteil „Leseverstehen" des Zertifikats Deutsch als Fremdsprache und einem deutschen C-Test, daß die beiden Tests zu etwa 25% die gleiche Eigenschaft erfassen (der Korrelationskoeffizient von 0,5 ist zu quadrieren). Da die kriterienbezogene Validität empirisch anhand der Testergebnisse ermittelt wird, sprechen viele Autoren auch von **empirischer Validität**.

Zur Einschätzung der empirischen Validität ist es wichtig zu wissen, daß die Objektivität einen Einfluß auf die Reliabilität hat und daß die Reliabilität wiederum eine Obergrenze für die empirische Validität eines Tests darstellt. Dies bedeutet, daß ein wenig objektiver und wenig reliabler Test nicht gleichzeitig valide sein kann. Dieser wichtige Sachverhalt wird von Praktikern häufig übersehen (vgl. jedoch Moss 1992, 1994). Umgekehrt bedeutet eine hohe Objektivität und Reliabilität keineswegs, daß der entsprechende Test auch valide ist, d.h. das erfaßt, was er erfassen soll. Ein bekanntes Schema zur Verdeutlichung der Beziehungen zwischen den Gütekriterien Objektivität, Reliabilität und Validität stammt von Lienert und ist mit geringfügigen Veränderungen in Abbildung 1 wiedergegeben.

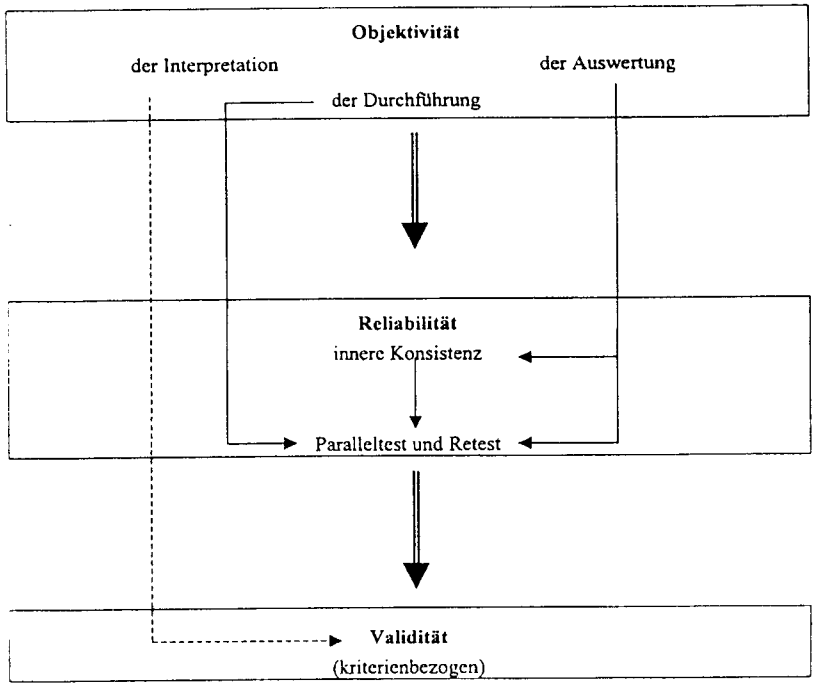

Abb. 1: Beziehungen zwischen den Gütekriterien Objektivität, Reliabilität und Validität (nach Lienert/Raatz 1994: 13)

Vor dem Hintergrund der vorangehenden Ausführungen ist die Abbildung 1 folgendermaßen zu interpretieren (vgl. Lienert/Raatz 1994: 13f.):

- Objektivität und Reliabilität sind notwendige, nicht jedoch hinreichende Voraussetzungen für eine zufriedenstellende Validität.
- Ein Test kann – kriterienbezogen – nicht valider als reliabel sein.
- Sowohl die Paralleltestreliabilität als auch die Retestreliabilität können nicht höher sein als die Konsistenz sowie die Auswertungs- und Durchführungsobjektivität.
- Eine hohe kriterienbezogene Validität impliziert zugleich eine hohe Objektivität und Reliabilität. Im Fall einer hohen kriterienbezogenen Validität kann deshalb häufig auf eine Überprüfung der Objektivität und Reliabilität verzichtet werden.
- Ein Test mit einer ausreichenden Validität und einer geringen Reliabilität hat ausgezeichnete Verbesserungschancen, da sich die Reliabilität – und damit zugleich die kriterienbezogene Validität – zumeist testtechnisch leicht erhöhen läßt (z.B. durch Aussondern und Hinzufügen von Aufgaben).
- Ein Test mit geringer Validität und hoher Reliabilität eignet sich zwar zur Differenzierung zwischen Individuen, jedoch nur sehr bedingt zur Vorhersage des jeweiligen Kriteriums (Test und Kriterium messen nur sehr bedingt das Gleiche). Die kriterienbezogene Validität eines solchen Tests kann nur über eine inhaltliche Überarbeitung verbessert werden.
- Um eine zufriedenstellende kriterienbezogene Validität zu erreichen, muß nicht nur der Test, sondern auch das Kriterium hinreichend objektiv und reliabel sein.

Ein weiterer Aspekt der Validität betrifft die Gültigkeit, die ein bestimmtes Verfahren in den Augen der Getesteten und der Testabnehmer hat. Diese sog. **Augenscheinvalidität** (engl.: *face validity*) ist nicht unwichtig für die Akzeptanz eines Verfahrens und damit auch z.B. für dessen prädiktive Gültigkeit. So haben häufig neue Testverfahren erst einmal eine geringe Augenscheingültigkeit für die angezielte Personengruppe (Lerner, Lehrende, Administratoren usw.). Gelingt es nicht, die Augenscheinvalidität z.B. durch gezielte Informationen über den Sinn des Tests zu erhöhen, kann dies den (praktischen) Wert des neu entwickelten Tests deutlich beeinträchtigen. So kann z.B. eine geringe Augenscheinvalidität dazu führen, daß die Kandidaten den Test nicht hinreichend ernst nehmen und deshalb nicht ihre optimale Leistung zeigen.

Bei der Bestimmung der **Konstruktvalidität** wird zunächst einmal gefragt, inwieweit die beobachteten Testergebnisse gültige Indikatoren von zugrundeliegenden theoretischen Konstrukten sind. So kann z.B. untersucht werden, inwieweit ein Multiple-Choice-Grammatiktest gültige Hinweise in bezug auf das theoretische Konstrukt „Kommunikative Kompetenz" liefert oder inwieweit ein Lesetest Aussagen zur Automatisierung von Leseverstehensprozessen ermöglicht. In die Bestimmung der Konstruktvalidität können eine Vielzahl von Überlegungen und Datenquellen eingehen – so z.B. (vgl. u.a. Grotjahn 1986, 1994a; Sigott 1994; Bachman/Eignor 1997; Kunnan 1998, 1999a, 1999b; Chapelle 1999):

- theoretische Überlegungen zur Konstruktspezifikation
- Korrelationen mit unterschiedlichsten Außenkriterien, d.h. Daten zur empirischen Validität
- quantitative und qualitative Daten zu den Prozessen bei der Testbearbeitung.

In jüngerer Zeit wird von einer Reihe von Autoren vor allem unter Bezug auf Messick (z.B. 1989, 1994 und 1996) die fundamentale Bedeutung der Konstruktvalidität als allen anderen Validitätsarten übergeordnetes Konzept herausgestellt. Hervorzuheben an der von Messick und anderen propagierten Konzeption ist vor allem, daß sich das Kriterium der Konstruktvalidität nicht nur auf die Gültigkeit der Operationalisierung des zu messenden Konstrukts, sondern auch auf den **Gebrauch der Testergebnisse** (z.B. in Form von Entscheidungen über einzelne Kandidaten) sowie auf die **Wertimplikationen** und **sozialen Konsequenzen** z.B. in Form eines positiven oder negativen "washback" auf den Unterricht bezieht (vgl. auch Messick 1996; Hamp-Lyons 1997; Wall 1997; McNamara 1998; Bailey 1999; Chapelle 1999; Kunnan 1999a, 1999b). Messick spricht in diesem Zusammenhang von der "consequential basis of test interpretation and test use" und stellt hierzu fest:

> "... evidence of the relevance and utility of test scores in specific applied settings, and evaluation of the social consequences of test use as well as of the value implications of test interpretation, all contribute in important ways to the construct validity of score meaning." (Messick 1989: 21)

Das von Messick und anderen vertretene weite Konzept der Konstruktvalidität hat erhebliche Implikationen hinsichtlich eines Tests wie dem TESTDAF, der politisch motiviert ist und weitreichende Konsequenzen für den einzelnen Testteilnehmer hat. Nicht umsonst beziehen sich Chapelle/Grabe/Berns (1997: 29ff.) im Rahmen der Entwicklung des in seiner Zielsetzung mit dem TESTDAF vergleichbaren TOEFL 2000[4] explizit auf das Konzept der Konstruktvalidität im Sinne von Messick.

Bei der Beurteilung der Konstruktvalidität eines Tests sind u.a. folgende Quellen möglicher Invalidität zu berücksichtigen: (a) Unterrepräsentation des zu messenden Konstrukts; (b) konstruktirrelevante Testvarianz (vgl. z.B. Messick 1989: 34). Im Fall einer **Unterrepräsentation des Konstrukts** ist der Test zu eng gefaßt und läßt wichtige Dimensionen des Konstrukts unberücksichtigt. Ein Extrembeispiel ist das Testen kommunikativer Kompe-tenz mit Hilfe eines reinen Wissenstests im Papier-und-Bleistift-Format. **Konstruktirrelevante Varianz** liegt vor, wenn bestimmte Merkmale, die nichts mit der zu messenden Fähigkeit zu tun haben, eine Aufgabe für bestimmte Probanden(gruppen) leichter oder schwerer machen. Dies ist z.B. der Fall, wenn man bei einem Lesetest den Faktor „thematisches Wissen" explizit aus der Spezifikation des zu messenden Konstrukts ausschließt und dann feststellt, daß das thematische Wissen dennoch einen differentiellen Effekt hat – z.B. bei Studierenden verschiedener Fachrichtungen (vgl. hierzu auch Grotjahn 2000).

[4] Es handelt sich hierbei um den Nachfolgetest des entsprechenden weltweit eingesetzten Tests.

Neben den aufgeführten Kriterien wird als zusätzliches Gütekriterium u.a. die **Normierung** eines Tests genannt (vgl. Ingenkamp 1985: 44ff.). Damit ist die Eichung des Tests für eine bestimmte Zielpopulation (Eichpopulation) gemeint. Viele Autoren verwenden hierfür auch den Terminus „Standardisierung".[5] Bei der Zielpopulation kann es sich z.B. um alle Schüler einer bestimmten Klassenstufe oder auch um alle Studienanfänger in einem bestimmten akademischen Fach handeln. Die Eichung erfolgt mit Hilfe statistischer Verfahren anhand einer repräsentativen Stichprobe (Eichstichprobe) aus der Zielpopulation (Eichpopulation) und führt zu sog. Normwerten (Testnormen). Liegen für eine bestimmte Zielpopulation geltende Normwerte vor, dann ist es möglich, den individuellen Testwert eines Lerners relativ zu den Leistungen der Zielpopulation zu beurteilen und nicht nur z.B. relativ zu den Leistungen der anderen Testteilnehmer. Das Urteil ist damit zwar unabhängig von der Leistung der jeweiligen Testgruppe; es ist jedoch nicht zugleich auch populationsunabhängig im Sinne der probabilistischen Testtheorie (vgl. Abschnitt 3). Geht es jedoch z.B. lediglich darum, interindividuelle Unterschiede innerhalb einer Gruppe aufzudecken, spielt die Normierung keine Rolle. Da das Gütekriterium der Normierung unabhängig von den übrigen Gütekriterien ist, kann auch ein normierter Test wenig objektiv, wenig reliabel und wenig valide sein.

Entsprechend den unterschiedlichen Bedeutungen von „Standardisierung" wird auch der Terminus „**standardisierter Test**" in unterschiedlicher Bedeutung verwendet. Für viele Autoren ist die Normierung das entscheidende Merkmal eines standardisierten Tests. Für andere Autoren ist dagegen die adäquate Standardisierung der Durchführung, Auswertung und Interpretation die wichtigste Charakteristik eines standardisierten Tests.[6] Zudem wird zu-

[5] Diese Bedeutung von Standardisierung ist nicht zu verwechseln mit der Standardisierung der Durchführung, Auswertung und Interpretation von Tests.

[6] So argumentiert z.B. bei Cronbach (1984: 27): "Tests having norms are sometimes called 'standardized tests.' I am not using the word in that sense, because I wish to emphasize standardization **of procedure**. A test may have a table of norms even though its procedures are not clearly specified, and a test with well-standardized procedures may not have norms. Obviously,

meist gefordert, daß ein standardisierter Test weitere Standards und zwar insbesondere die Hauptgütekriterien in hinreichendem Maße erfüllt. Im übrigen ist eine Normierung nur dann sinnvoll, wenn sich die Hauptgütekriterien in empirischen Analysen als adäquat erwiesen haben.

Für **informelle Tests** ist u.a. kennzeichnend, daß auf eine Normierung an einer repräsentativen Stichprobe verzichtet wird. Da informelle Tests vor allem im Unterricht zur Aufdeckung interindividueller Unterschiede eingesetzt werden, ist das Fehlen einer Normierung kein Nachteil. Das wichtigste Gütekriterium eines informellen Tests ist dessen **Situationsvalidität** und dessen **Inhaltsvalidität** (vgl. Schwarzer/Schwarzer 1982: 318ff.).

Ein weiteres zuweilen genanntes Gütekriterium bezieht sich auf die **Authentizität** der Testaufgaben. In diesem Zusammenhang kann mit „authentisch" z.B. gemeint sein, daß es sich um genuine, nicht spezifisch für den Test produzierte Aufgaben handelt. Weiterhin kann sich „authentisch" auf den Grad der Übereinstimmung zwischen den Merkmalen einer gegebenen Testaufgabe und den Merkmalen der jeweiligen zielsprachlichen Aufgabe beziehen.[7] Authentizität im letztgenannten Sinne ist u.a. nach Bachman/Palmer (1996) eine wichtige Charakteristik der Konstruktvalidität und damit der Qualität eines Tests – und dies aus zumindest zwei Gründen: Authentische Testaufgaben erlauben zum einen Generalisierungen im Hinblick auf die Fähigkeit zur Lösung analoger zielsprachlicher Probleme außerhalb der Testsituation. Zum anderen bestimmt die Authentizität einer Testaufgabe die Wahrnehmung der Aufgabe durch den jeweiligen Probanden – die Aufgabe wird im Sinne der Augenscheinvalidität z.B. als relevant angesehen. Dies kann wiederum einen Einfluß auf die Testleistung haben (vgl. Bachman/Palmer 1996: 24 sowie auch McNamara 1996: Kap. 2; Douglas 1997: 116). Bachman/Palmer (1996: 24) stellen in diesem Zusammenhang u.a. fest: "It is this relevance, as perceived by the test taker, that we believe helps promote **a positive affective response** to the test task and can thus help test takers **to perform at their best**." (Hervorhebung R. G.)

collecting norms is not profitable until procedures are standardized." (Hervorhebung im Original)

[7] So bei Bachman/Palmer (1996: 23). Vgl. auch Edelhoff (1985); Spolsky (1985); Peirce (1992: 681f.); Doyé (1993); Decoo/Colpaert (1997: 31ff.); Lewkowicz (1997a, 1997b); Glaboniat (1998: 70); Amor (1999).

Wie allerdings Lewkowicz (1997a, 1997b) zu Recht feststellt, gibt es bisher kaum empirische Belege für die Bedeutung des Faktors „Authentizität". Die Autorin gelangt aufgrund ihrer eigenen empirischen Untersuchungen zu Hörverstehensaufgaben u.a. zu folgenden Feststellungen: (1) Es ist durchaus möglich und auch akzeptabel, „authentisch **aussehende**" pädagogische Texte im Rahmen von Tests einzusetzen. (2) Nur für wenige Testteilnehmer ist die Authentizität einer Testaufgabe ein wichtiges Merkmal (1997a: 182). Im übrigen hat bereits Spolsky (1985: 39) darauf hingewiesen, daß Sprachtesten notwendigerweise Inauthentizität impliziert und dies treffend folgendermaßen formuliert: "Any language test is by its very nature inauthentic, abnormal language behaviour ..."

Entsprechend steht auch bei der Entwicklung des TOEFL 2000 nicht in erster Linie die Authentizität, sondern vor allem die Konstruktvalidität der Testaufgaben im Vordergrund. Chapelle/Grabe/Berns (1997: 26) stellen in diesem Zusammenhang fest:

> "Because test situations are inherently different from the contexts about which we want to infer test takers' ability, students' performance on a test is likely to offer a distorted picture of the ability they would use in 'authentic' contexts. The issue is, then, how we can use the picture of ability obtained by test performance to make inferences about abilities in other contexts. In order for tests to be used appropriately, it is the responsibility of test developers to demonstrate, and test users to consider, evidence concerning the meaning of test scores (i.e., construct validity evidence). To investigate construct validity, it is necessary to hypothesize the construct that the test is intended to measure. Developing a test whose validity can be justified is the primary objective of TOEFL 2000 ..."

Auch in bezug auf den TESTDAF ist wiederholt gefordert worden, daß die Testaufgaben möglichst authentisch gestaltet werden sollten. In Anbetracht der vorangehenden Ausführungen sollte m.E. das Kriterium der Authentizität nicht zu sehr in den Vordergrund gestellt werden. Eine hohe Authentizität dürfte zwar zur Erhöhung der Akzeptabilität des TESTDAF bei Testadministratoren und Lehrenden beitragen, sie muß allerdings nicht gleichzeitig auch zu valideren Testergebnissen führen (vgl. auch den Überblick zum "washback effect" bei Wall 1997 und Bailey 1999). Im Fall eines für die Teilnehmer persönlich wichtigen Tests wie dem TESTDAF ist vielmehr zu erwarten, daß die Probanden versuchen, auch Aufgaben, die von ihnen als wenig authentisch wahrgenommen werden, möglichst optimal zu lösen. Wie beim TOEFL 2000

steht deshalb zu Recht beim TESTDAF die Sicherung der Konstruktvalidität im Vordergrund.

Wichtiger als die von den Testteilnehmern wahrgenommene Authentizität dürfte die **tatsächliche Vertrautheit** der Probanden mit bestimmten Aufgabenformen sein. Da Unterschiede in der Vertrautheit mit den verwendeten Formaten einen differentiellen Effekt auf die Aufgabenschwierigkeit haben können und dies wiederum die Validität eines Tests beeinträchtigen kann, ist ein entsprechender Effekt möglichst weitgehend auszuschließen. Dies kann z.B. dadurch geschehen, daß den Probanden die Möglichkeit gegeben wird, sich gründlich auf den Test vorzubereiten.

Ich hatte bereits darauf hingewiesen, daß für eine Reihe von Autoren die **Nützlichkeit** eines Tests ein zentrales Gütekriterium ist. So heißt es z.B. bei Bachman/Palmer (1996: 17):

> "The most important consideration in designing and developing a language test is the use for which it is intended, so that **the most impor-tant quality of a test is its usefulness**." (Hervorhebung R.G.)

Damit ist für Bachman/Palmer (1996) die Nützlichkeit sogar das letztendlich entscheidende Kriterium für die Güte eines Tests. Nicht umsonst lautet der Titel ihres Buches: "Language Testing in Practice: Designing and Developing Useful Language Tests."

Bachman/Palmer (1996: 18) definieren Nützlichkeit als Funktion von insgesamt sechs komplementären Eigenschaften:

Usefulness =
Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality[8]

[8] Diese Definition unterscheidet sich grundlegend von dem Verständnis der Nützlichkeit als Nebengütekriterien bei Lienert/Raatz (1994: 13): „Ein Test ist dann **nützlich**, wenn er ein Persönlichkeitsmerkmal oder Verhaltensweise mißt oder vorhersagt, für dessen Untersuchung ein praktisches Bedürfnis besteht. Ein Test hat demgemäß eine hohe Nützlichkeit, wenn er in seiner Funktion durch keinen anderen Test vertreten werden kann, und er hat eine geringe Nützlichkeit, wenn er ein Persönlichkeitsmerkmal prüft, das mit einer Reihe anderer Tests ebenso gut untersucht werden könnte."

Auf die ersten drei Eigenschaften bin ich bereits eingegangen. Nach Bachman/Palmer (1996: 25f.) bezieht sich "interactiveness" auf das Ausmaß und Art der Wechselwirkung zwischen den Testaufgaben und den im Hinblick auf das zu messende Konstrukt relevanten kognitiven Merkmalen des jeweiligen Kandidaten (insbesondere den Wissensbeständen, metakognitiven Strategien und affektiven Schemata). Testaufgaben können in unterschiedlichem Grade das Merkmal der Interaktivität aufweisen. Will man z.B. die mündliche Kommunikationsfähigkeit testen, dann hat ein mündliches Interview, in dem der Kandidat den Gesprächsverlauf mitbestimmen kann, eine höhere Interaktivität als z.B. ein simuliertes mündliches Interview, in dem die Gesprächsstimuli in starrer Reihenfolge über einen Tonträger präsentiert werden.[9] Unter "impact" verstehen die Autoren den Einfluß des Tests sowohl auf den einzelnen Testkandidaten als auch z.B. auf das jeweilige Erziehungssystem oder die jeweilige Gesellschaft (andere Autoren sprechen hier von "washback" oder auch "backwash"). Schließlich definieren sie die Praktikabilität eines Tests als das Verhältnis von vorhandenen Ressourcen zu notwendigen Ressourcen.

Bachman/Palmer (1996: 18) nennen drei zentrale Prinzipien, die es bei der Sicherung der Nützlichkeit eines Tests zu beachten gilt:

Prinzip 1:	Die Gesamtnützlichkeit eines Tests ist zu maximieren und nicht einzelne Komponenten der Nützlichkeit wie z.B. die Reliabilität.
Prinzip 2:	Die Komponenten der Nützlichkeit können nicht unabhängig voneinander bewertet werden, sondern lediglich hinsichtlich ihrer kombinierten Wirkung auf die Nützlichkeit des Tests.
Prinzip 3:	Der Grad der Nützlichkeit und die angemessene Balance zwischen den einzelnen Komponenten kann nicht allgemein festgelegt werden, sondern muß jeweils in Abhängigkeit von der spezifischen Testsituation bestimmt werden.

Sie illustrieren die Bedeutung der Prinzipien, und zwar insbesondere des Prinzips 3, u.a. an folgendem treffenden Beispiel:

[9] Ein deutlich weiteres, soziolinguistisch orientiertes Konzept von Interaktivität vertritt McNamara (1997a).

> "In a large-scale test that will be used for making important decisions about large numbers of individuals, for example, the test developer may want to design the test and test tasks so as to achieve the highest possible levels of reliability and validity. In a classroom test, on the other hand, the teacher may want to utilize test tasks that will provide higher degrees of authenticity, interactiveness and impact." (Bachman/Palmer 1996: 19)

Die Autoren haben m.E. mit dem explizierten Konzept von Nützlichkeit ein sehr sinnvolles Kriterium zur Bewertung von Tests vorgelegt. Dies gilt insbesondere in Bezug auf die Praxis. Denn gerade einem Praktiker dürfte ein pragmatisch fundiertes Qualitätskonzept, das auch die Rückwirkung des Tests auf den Unterricht und die Praktikabilität des Verfahrens berücksichtigt, unmittelbar einleuchten.

5. Performanztests vs. Kompetenztests

Sprachtests lassen sich u.a. anhand der folgenden beiden voneinander unabhängigen Merkmalsdimensionen beschreiben:

1. Testaufgaben und vom Kandidaten im Test erwartete Reaktionen modellieren (replizieren) zielsprachliche Aufgaben und Verwendungssituationen (d.h. Aufgaben und Verwendungssituationen außerhalb des Testkontexts).
2. Den Testaufgaben liegt ein (explizites) Modell der beim Gebrauch von Sprache involvierten Fähigkeiten und Fertigkeiten zugrunde.

Beide Merkmalsdimensionen bilden jeweils eine Rangskala, d.h., das Merkmal kann auf eine Testaufgabe und damit auch auf einen Test in unterschiedlichem Maße zutreffen. Modellieren die Testaufgaben und die im Test erwarteten Reaktionen die zielsprachlichen Aufgaben und Verwendungssituationen (das Kriterium), dann wird ein entsprechender Test auch als **Performanztest**[10] bezeichnet. Die wesentlichen Merkmale eines Performanztests sind in Abbildung 2 dargestellt (aus McNamara 1997a: 448).

[10] Der Terminus „Performanztest" wird in sehr unterschiedlicher Weise verwendet. Eine Reihe von Definitionen werden in McNamara (1996: Kap. 2) diskutiert. Will man den nicht unproblematischen Anglizismus „Performanz" vermeiden, dann kommen als Äquivalente u.a. „Sprachgebrauchstest" oder „Sprachverwendungstest" in Betracht.

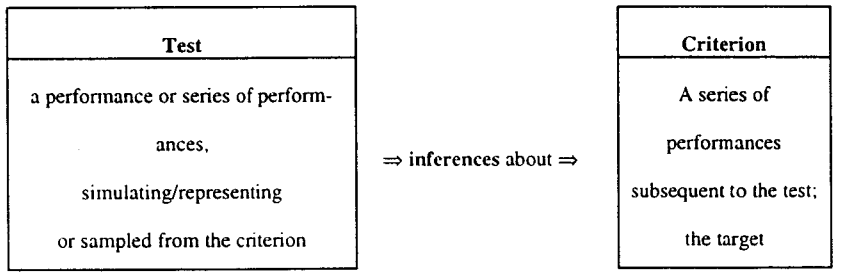

Abb. 2: Merkmale eines Performanztests

Wie die Abbildung 2 zeigt, werden anhand der im Test gezeigten Performanz Schlußfolgerungen gezogen hinsichtlich zukünftiger Performanz des Kandidaten außerhalb der Testsituation z.B. als Arzt, Fremdenführer, Kellner oder Studierender einer bestimmten Fachrichtung. Die beobachtete Performanz der Kandidaten wird häufig anhand von Ratingskalen beurteilt. Dies ist z.B. in den Testteilen „mündlicher Ausdruck" und „schriftlicher Ausdruck" des TESTDAF der Fall. Die Güte der Schlußfolgerungen hängt damit u.a. auch von der Qualität der Ratingskalen und von der Qualität der Beurteiler ab (diese können z.B. zu milde oder zu streng oder auch inkonsistent urteilen).

Einem Performanztest liegen u.a. folgende Annahmen zugrunde: (1) Fähigkeiten wie z.B. grammatikalische oder lexikalische Kompetenz sind zwar bis zu einem gewissen Grad eine Vorbedingung für einen effektiven Sprachgebrauch, sie können diesen jedoch keineswegs garantieren. (2) Um tatsächliches sprachliches Verhalten vorherzusagen, ist es nötig, Performanzfaktoren wie akustische Bedingungen, kommunikatives Ziel oder Verhältnis der Kommunikationsteilnehmer zueinander explizit bei der Konstruktion der Testaufgaben zu berücksichtigen. (3) Je genauer die Testaufgaben ein bestimmtes zielsprachliches Verhalten außerhalb der Testsituation abbilden, desto präziser kann dieses Verhalten vorhergesagt werden (vgl. auch Skehan 1998: 153ff.).

Mit McNamara (1996: 43ff.) lassen sich zwei Typen von Performanztests unterscheiden: Performanztests in einem **starken** und in einem **schwachen** Sinne.[11] Im Fall eines Performanztests im starken Sinne duplizieren die Testaufgaben und die vom Kandidaten erwarteten Reaktionen so weit wie möglich die vom Testkandidaten außerhalb der Testsituation erwarteten Verhaltensweisen, und das letztendlich entscheidende Erfolgskriterium ist nicht

[11] McNamara (1996: 43) weist darauf hin, daß es sich bei der Unterscheidung von "strong and weak language performance tests" lediglich um eine analytische Trennung handelt.

die vom Kandidaten gezeigte sprachliche Leistung, sondern die Lösung des gestellten Problems. Manche Autoren sprechen in diesem Zusammenhang auch von **authentischen Tests**. Ist der Kandidat z.B. ein angehender Arzt und besteht die Testaufgabe in der Erhebung von anamnestischer Information bei einem Patienten, dann ist das entscheidende Erfolgskriterium die Qualität der erhaltenen Informationen und nicht etwa Umfang, Komplexität und Korrektheit der Äußerungen des Kandidaten (vgl. McNamara 1996: 82f. sowie auch Skehan 1998: 181). Ein Performanztest im starken Sinne ist zugleich auch ein sog. **direkter Test**, da von der im Test gezeigten Leistung (relativ) direkt auf die Leistung des Kandidaten außerhalb der Testsituation geschlossen wird. Wegen ihrer Situationsspezifität haben Performanztests im starken Sinne nur einen sehr eingeschränkten Einsatzbereich. Für diesen haben sie allerdings eine hohe Vorhersagevalidität.

Eine Vielzahl von Beispielen für Aufgaben, die das Kriterium eines starken Performanztests mehr oder minder erfüllen, findet sich in Norris/Brown/Hudson/Yoshioka (1998). Entsprechende Aufgaben sollten nach den genannten Autoren folgende Eigenschaften aufweisen:

> "The tasks should:
> a) Be based on needs analysis (including student input) in terms of rating criteria, content, and contexts
> b) Be as authentic as possible with the goal of measuring real-world act ivities
> c) Sometimes have collaborative elements that stimulate communicative interactions
> d) Be contextualized and complex
> e) Integrate skills with content
> f) Be appropriate in terms of number, timing, and frequency of assessment
> g) Be generally non-intrusive, that is, be aligned with the daily actions in the language classroom" (Norris/Brown/Hudson/Yoshioka 1998: 10f.)

Bei einem Performanztest im schwachen Sinne steht dagegen die **sprachliche** Performanz im Vordergrund. Die Testaufgaben selbst müssen die zielsprachlichen Verwendungssituationen auch nicht so weit wie möglich duplizieren, sondern lediglich die als wesentlich erachteten Merkmale abbilden. Das entscheidende Kriterium ist nicht die Erfüllung der gestellten Aufgabe, sondern die im Test gezeigte sprachliche Leistung. Ein Beispiel für einen Performanztest im schwachen Sinne ist der Testteil „Mündlicher Ausdruck" im TESTDAF. Es handelt sich hierbei um ein "Simulated Oral Proficiency Interview" (SOPI), bei dem die Stimuli per Tonträger präsentiert und die Antworten der Testpersonen auf Tonträger aufgenommen werden. Ein weiteres Beispiel ist das *Oral Proficiency Interview* (OPI) des *American Council on the Teaching of Foreign Languages* (ACTFL). Das OPI ist zwar ein face-to-face-Interview mit einer im Vergleich zum SOPI größeren Wirklichkeitsnähe; das entscheidende Kriterium ist jedoch wie beim SOPI die vom Kandidaten gezeigte sprachliche Leistung. Bei einem Performanztests im schwachen Sinne ist die Beziehung zwischen Test und Kriterium deutlich weniger direkt als bei einem Performanztest im starken Sinne. Sie können deshalb auch zu den **semidirekten Tests** gezählt werden.

McNamara (z.B. 1996, 1997b) unterscheidet neben Performanztests im starken Sinne und im schwachen Sinne zwei unterschiedliche Traditionen bzw. Ansätze im Bereich des Performanztestens: "the work sample tradition" und "the cognitive and psycholinguistic tradition". Er charakterisiert diese folgendermaßen:

> "The two traditions of second language performance assessment are rather separate, although there is considerable blurring of the lines: the **work sample** tradition derives largely from work on performance assessment in non-language settings, particularly in personnel selection ... This approach ... is clearly illustrated in language testing for academic and specific occupational purposes, but it had also had an important influence on general purpose performance testing, in particular the important traditions of Oral Proficiency Interview testing now associated with the 'Proficiency Movement' ... The key to this approach is the (more or less) realistic representation of relevant real world tasks in the test setting. ... The focus ... is on careful specification of the content of the test in terms of the criterion ... A job analysis of the target situation is carried out, often informed by a sociolinguistic framework for analysing aspects of the target language use setting. In this approach, the performance is the **target** of assessment (Messick, 1994). Considered from a theoretical point of view, this tradition can be characterized as behaviour-based and sociolinguistic in orientation. ... In the second, more general, cognitive and psycholinguistic tradition of performance testing, the performance task itself is of less interest than what the performance reveals of underlying ability. In the words of Messick (1994), the performance is the **vehicle** of assessment, rather than itself

being the **target**." (McNamara 1997b: 132f.; Hervorhebungen im Original)

McNamaras Differenzierung zwischen einem berufsbezogenen und einem kognitiv-psycholinguistischen Ansatz weist zwar deutliche Affinitäten zu der Unterscheidung zwischen Performanztests im starken und im schwachen Sinne auf, sie ist jedoch keineswegs damit deckungsgleich. Ein typisches Beispiel für ein Auseinanderklaffen der beiden Differenzierungsdimensionen ist ein mündliches Interview auf der Basis des "work sample approach" mit einem zugleich eindeutigen Fokus auf der **sprachlichen** Leistung (Performanztest im schwachen Sinne).

Performanztests lassen sich weiterhin danach beurteilen, inwieweit den Testaufgaben ein (explizites) Modell der beim Gebrauch von Sprache involvierten Fähigkeiten und Fertigkeiten zugrunde liegt. Je indirekter ein Performanztest die Kriteriumsleistung mißt, desto dringender benötigen wir Modelle der Fähigkeiten und Fertigkeiten, die zum einen der Lösung der Testaufgaben und zum anderen dem Sprachgebrauch außerhalb der Testsituation zugrunde liegen. Erst der Bezug auf entsprechende empirisch validierte Modelle ermöglicht es, z.B. kontextspezifische und kontextunspezifische Merkmale von Testaufgaben zu unterscheiden. Dies erlaubt wiederum eine bessere Beurteilung der Generalisierbarkeit von Testresultaten oder auch die Identifikation sog. Methodeneffekte. Autoren wie McNamara (1996) und Skehan (1998: Kap. 7) sprechen sich deshalb zu Recht für eine stärkere psycholinguistische und auch soziolinguistische Fundierung von Performanztests aus.[11]

In bezug auf die mangelnde theoretische Fundierung vieler Performanztests und zwar insbesondere solcher der "work sample tradition" stellt McNamara (1997b) u.a. fest:

> "In judging test performances ... we are not interested in the observed performances for their own sake; if we were, and that is all we are interested in, the sample performance would not be a test. Instead we are more interested in what the performance reveals of the potential for subsequent performances in the criterion situation; we seek in the test performance those qualities which are indicative of what is held to underlie it. For this **we need a theory (implicit or explicit) of the relation between test performance and criterion behaviour. The relative lack of such a theoretical grounding of practice is a weakness in much performance testing, which tends to be resolutely atheoretical**, with the result that the bases for inferences about test candidates remain unclear, and the threats to the validity of those inferences not open to investigation." (McNamara 1997b: 133; Hervorhebungen R. G.)

[11] Bei Skehan (1998) findet sich eine interessante Diskussion von psycholinguistischen Aspekten von Performanztests. McNamara (1996) geht umfassend auch auf spezifische Meßprobleme ein. Soziolinguistische Aspekte von mündlichem Sprachgebrauch in und außerhalb von Testsituationen werden u.a. in Young/He (1998) thematisiert.

Wie bereits angedeutet haben Performanztests sowohl Vorteile als auch Nachteile. Ein Vorteil insbesondere von Performanztests im starken Sinne ist die relativ große Wirklichkeitsnähe und Authentizität und damit verbunden die hohe Augenscheinvalidität sowie der potentiell positive "washback" (Auswirkungen auf Unterricht, Curriculum usw.). Dies gilt vor allem dann, wenn Sprache für genau spezifizierte berufsbezogene Aufgaben eingesetzt werden soll, wie z.B. Deutsch als Fremdsprache im Hotelgewerbe, oder auch im Fall eines kommunikativ ausgerichteten Fremdsprachenunterrichts (vgl. auch die Diskussion des Gütekriteriums der Authentizität weiter oben).

Ein Nachteil von Performanztests vor allem im starken Sinne ist die Subjektivität der Auswertung, die häufig nicht zufriedenstellende Reliabilität sowie die (relativ) geringe situations- und aufgabenübergreifende Generalisierbarkeit der Testresultate. Insbesondere die geringe situations- und aufgabenübergreifende Generalisierbarkeit der Testresultate ist für viele Testspezialisten eine entscheidende Schwäche starker Performanztests, da sie den Anwendungsbereich massiv einschränkt.

Weiterhin wird zuweilen auch zwischen linguistischen und kommunikativen Performanztests differenziert (vgl. Klein-Braley 1991; Grotjahn 1994b).

In **linguistischen Performanztests** werden Stichproben kontextueller Sprachverwendung erhoben, ohne daß jedoch den Probanden eine kommunikative Aufgabe, wie das Führen einer Beschwerde, gestellt wird. Ein mögliches Beispiel für einen linguistischen Performanztest ist der Cloze Test, der von McNamara (1996: 79f.; 1997b: 133) unter Bezug auf Oller (1979) der kognitiv-psycholinguistischen Tradition des Performanztestens zugerechnet wird.

In **kommunikativen Performanztests** werden den Lernern Aufgaben gestellt, in denen sie für das Alltags- oder Berufsleben typische Ziele mit Hilfe von Sprache strategisch realisieren sollen. Ein Beispiel für eine Aufgabe in einem kommunikativen Performanztest ist das Schreiben eines Bewerbungsbriefes.

Die Bewertung des sprachlichen Produkts erfolgt fast stets **global** auf der Basis von Schätzskalen. Kommunikative Performanztests fallen damit unter die Rubrik „Performanztests im starken Sinne".

Bei der Beurteilung des Stellenwerts kommunikativer Performanztests ist u.a. zu berücksichtigen, daß die kommunikativ zu realisierenden Ziele nicht aufgrund persönlicher Bedürfnisse der Probanden verfolgt werden, sondern durch den Test gesetzt sind. Folglich handelt es sich auch nur um simulierte Kommunikation. Auch sog. kommunikative Tests sind deshalb niemals **wirklich** kommunikativ.

Kompetenztests[12] stehen im Gegensatz insbesondere zu Performanztests im starken Sinne. Sie beruhen u.a. auf folgenden Annahmen:

1. Sprachliches Verhalten außerhalb der Testsituation kann auf zugrundeliegende Fähigkeiten zurückgeführt werden.
2. Die zugrundeliegenden Fähigkeiten bestehen aus unterschiedlichen, miteinander in Wechselwirkung stehenden Komponenten.
3. Sprachtests, die die zugrundeliegenden Fähigkeiten valide erfassen, erlauben generalisierende Aussagen über sprachliches Verhalten außerhalb der Testsituation.
4. Je umfassender das dem Test zugrundeliegende Modell sprachlicher Kommunikationsfähigkeiten ist und je umfassender der Test dieses repräsentiert, desto größer ist die Zahl potentieller sprachlicher Verwendungssituationen, für die der Test Vorhersagen ermöglicht (vgl. auch Skehan 1998: 153ff.).

Kompetenztests sind damit zumindest prinzipiell für eine Vielzahl unterschiedlicher Verwendungssituationen relevant. Dies ist ein möglicher Vorteil vor allem im Vergleich zu eindeutig berufsbezogenen Performanztests. Dem steht allerdings als Nachteil die potentiell geringere Validität und der nicht auszuschließende negative Rückwirkungseffekt gegenüber.

Ein Extremfall eines sprachlichen Kompetenztests ist ein **Sprachwissenstest**.[13] Mit diesem Terminus sollen Tests bezeichnet werden, die in erster Linie

[12] Der Begriff „Kompetenz" ist hier nicht mit dem Kompetenzbegriff von Chomsky gleichzusetzen, sondern umfasst die „Fähigkeit zum Gebrauch von Sprache in realen Verwendungssituationen" (vgl. die Diskussion weiter unten).

sprachliches Wissen ohne auch nur eine indirekte Berücksichtigung von Performanzfaktoren überprüfen. Hierunter fallen vor allem die sog. discrete-point-Tests, die spezifische sprachliche Wissensbestände – wie z.B. die Kenntnis des Konjunktivs im Deutschen – in (weitgehend) dekontextualisierter Form testen.

Sprachwissenstests im definierten Sinne messen damit in erster Linie deklaratives Wissen, d.h. sog. „Wissen, daß etwas der Fall ist", nicht jedoch prozedurales Wissen, d.h. „Wissen, wie man etwas tut" (vgl. zu dieser kognitionspsychologischen Unterscheidung z.B. Grotjahn 1994a, 1997 sowie Davies 1989).

Vor dem Hintergrund der getroffenen Unterscheidungen sind z.B. die Testteile „Leseverstehen" und „Hörverstehen" des TESTDAF eher dem Typ „Kompetenztest" zuzurechnen. Im Fall der Testteile „schriftlicher Ausdruck" und „mündlicher Ausdruck" handelt es sich dagegen eindeutig um Performanztests – allerdings eher im schwachen, kognitiv-psycholinguistischen Sinne.[14] Im Abschnitt 9 werde ich kurz darauf eingehen, warum im Fall des Testteils „mündlicher Ausdruck" ein Performanztest im schwachen Sinne gewählt worden ist.

6. Normorientierte vs. kriteriumsorientierte Tests

Grundlegend ist auch die Unterscheidung zwischen normorientierten (auch bezugsgruppenorientierten) und kriteriumsorientierten Tests. Das primäre Kriterium für diese Differenzierung ist zunächst einmal die Art der Interpretation der Testergebnisse. Man spricht deshalb auch von einer normorientierten und von einer kriteriumsorientierten Interpretation von Testergebnissen. Zusätzlich können sich die beiden Testformen jedoch z.B. auch in der Art der Testentwicklung, in der Teststruktur, in den Testinhalten und in den Testzielen unterscheiden.
Bei den **normorientierten Tests** werden die individuellen Ergebnisse **relativ** zu den Ergebnissen einer Bezugsgruppe interpretiert. Bei der Bezugsgruppe kann es sich z.B. um Mitschüler oder auch um eine repräsentative Vergleichs-

[13] Zuweilen wird „Sprachwissenstest" auch als Synonym zu „Kompetenztest" verwendet. „Sprachwissenstest" steht dann in Opposition zu „Sprachgebrauchstest" bzw. „Sprachverwendungstest".

[14] Die Testteile sind dem kognitiv-psycholinguistischen Ansatz zuzurechnen, weil in die Aufgabenkonstruktion explizite Hypothesen über kognitiv-psycholinguistische Verarbeitungsprozesse eingehen (z.B. in Form von Annahmen über den kognitiven Verarbeitungsaufwand).

gruppe (Eichstichprobe) handeln. Die Ergebnisse werden häufig z.B. folgendermaßen formuliert: Der Kandidat gehört zu den oberen 10% der Gruppe. Wie viele Aufgaben der Kandidat gelöst hat, bleibt hierbei unerwähnt. Ziel ist eine möglichst verläßliche Differenzierung zwischen den Kandidaten. Im Fall normorientierter Tests kann den Kandidaten zwar das Testformat durchaus bekannt sein; zumeist wissen sie jedoch nicht, welche spezifischen Kenntnisse und Fertigkeiten durch die Aufgaben getestet werden sollen.

Mit Hilfe von **kriteriumsorientierten Tests** will man dagegen ermitteln, ob und eventuell auch in welchem Ausmaß ein Lernender ein im Detail beschriebenes Kriterium, wie z.B. Kommunikationsfähigkeit als Fluglotse im Englischen oder ein bestimmtes Ausmaß an Deutschkenntnissen für das Hotelgewerbe, erreicht hat.[9] Die Testaufgaben repräsentieren das Kriterium, und das Ziel ist, den individuellen Fähigkeitsgrad eines Lernenden mit einem gewünschten Fähigkeitsgrad zu vergleichen (vgl. Ingenkamp 1985: 118, 124 sowie Klauer, 1987). Da die Leistung der übrigen Testteilnehmer dabei keine Rolle spielt, ist die Bewertung bezogen auf die Gruppe **absolut**. Handelt es sich beim Kriterium um Lehr- bzw. Lernziele, wird auch spezifischer von lehrzielorientierten bzw. lernzielorientierten Tests gesprochen. Im Fall lehrzielorientierter Tests wissen die Kandidaten in der Regel relativ genau, welche spezifischen Inhalte und Fertigkeiten durch den Test erfasst werden sollen. Lehrzielorientierte Tests zeichnen sich aufgrund ihres Unterrichtsbezugs meist durch eine höhere Augenscheingültigkeit und durch eine positive Rückwirkung auf den Unterricht aus. Sie sind deshalb, sofern eine Ausrichtung des Testens an Lernzielen möglich und sinnvoll ist, normorientierten Verfahren vorzuziehen (vgl. hierzu auch Hughes 1989 sowie Brown 1996: 1ff.).

Während der TOEFL und auch der TOEFL 2000 eher Beispiele für einen normorientierten Test darstellen, ist die Deutsche Sprachprüfung für den Hochschulzugang (DSH) hinsichtlich der Interpretation der Testergebnisse eindeutig ein kriteriumsorientierter Test. Zugrunde gelegt wird das Kriterium „ausreichende Deutschkenntnisse für die Aufnahme des Fachstudiums", und es wird lediglich das Urteil „bestanden/nicht bestanden" vergeben – und zwar unabhängig davon, welche Leistung der Kandidat im Vergleich zu den übrigen Testteilnehmern erbracht hat (vgl. auch Lee 1998: 9f.). Auch beim TESTDAF ist eine entsprechende kriterienorientierte Interpretation der Testergebnisse

[9] Diese Bedeutung des Begriffs ‚Kriterium' ist nicht zu verwechseln mit der Verwendungsweise des Begriffs im Sinne eines Beurteilungsmaßstabes für die empirische Validität eines Tests (vgl. Abschnitt 4).

vorgesehen. Dies ist neben Unterschieden im Testformat und in den Inhalten ein weiterer Grund, warum der TESTDAF keineswegs als „kleiner Bruder des TOEFL" (Ruhr-Nachrichten vom 11.8.98) anzusehen ist.

7. Einsatzmöglichkeiten von Sprachtests

Sprachtests werden vor allem dazu eingesetzt, um Entscheidungen zu treffen. Die Entscheidungen können sich auf einzelne Individuen, Gruppen von Individuen oder auch auf ganze Ausbildungsprogramme beziehen (vgl. zum Folgenden Bachman 1990: 58ff.; Grotjahn 1994b: 37f.; Grotjahn/Klein-Braley 1998: 297f.).

Welcher Testtyp in einer bestimmten Situation verwendet werden sollte, hängt dabei zunächst einmal von dem speziellen Ziel ab, mit dem der Einsatz verbunden ist. Zusätzlich können jedoch auch z.B. die verfügbaren Ressourcen eine entscheidende Rolle spielen. Ist man z.B. an spezifischer diagnostischer Information hinsichtlich der (expliziten) Kenntnis bestimmter Grammatikregeln interessiert, dann ist in vielen Fällen ein Kompetenztest in Form von Multiple-Choice-Aufgaben ein adäquates Verfahren. Ein entsprechender Test kann aber auch dann indiziert sein, wenn man darüber hinaus auch an der Fähigkeit zur Verwendung des jeweiligen Wissens interessiert ist. Dies gilt vor allem dann an, wenn folgende Bedingungen erfüllt sind:

1. Eine große Zahl von Probanden, da Multiple-Choice-Aufgaben maschinell auswertbar sind.
2. Das Vorliegen eines balancierten Fremdsprachenunterricht, d.h. eines weder skill- noch komponentenspezifischen Sprachunterrichts. Denn in einem solchen Kontext ist ein relativ starker Zusammenhang zwischen den sprachlichen Wissensbeständen und der korrekten Verwendung der entsprechenden Wissensbestände zu erwarten.

Bereitet dagegen ein Kurs auf die sprachliche Bewältigung bestimmter fest umrissener beruflicher Situationen vor, dann dürfte ein kommunikativer Performanztest die erste Wahl sein: Er hat in der Regel eine im Vergleich zu einem reinen Kompetenztest höhere Augenscheinvalidität und zudem einen potentiell positiven Effekt auf den Unterricht.

8. Modellierung sprachlicher Fähigkeiten und Fertigkeiten

Ich hatte bereits auf folgenden grundlegenden Sachverhalt hingewiesen: Je indirekter ein Test die Kriteriumsleistung mißt, desto dringender benötigen wir Modelle der Fähigkeiten und Fertigkeiten, die dem Sprachgebrauch innerhalb und außerhalb der Testsituation zugrunde liegen. In diesem Zusammenhang

stellt sich allerdings das Problem, daß es eine Vielzahl unterschiedlicher Modelle sprachlicher Fähigkeiten und Fertigkeiten und erheblich differierende Begrifflichkeiten gibt. So werden neben dem Begriff „kommunikative Kompetenz" (siehe z.B. Hymes 1971, 1972 sowie die Belege in Davies 1989) Begriffe wie "communicative language proficiency" (Chapelle/Grabe/Berns 1997) oder auch "communicative language ability" (Bachman 1990) verwendet. Kennzeichnend ist, daß die genannten Begriffe in der Regel nicht nur Sprachwissen im Sinne des abstrakten Kompetenzbegriffs von Chomsky (1965), sondern darüber hinaus auch die **Fähigkeit zu aktuellem Sprachgebrauch in konkreten Situationen** umfassen (Hymes spricht in diesem Zusammenhang von "ability for use").

Im Rahmen dieses Beitrags werde ich mich auf die Darstellung eines Modells beschränken. Weitere Hinweise finden sich z.B. in Taylor (1988); Davies (1989); Spolsky (1989); Widdowson (1989); Bachman (1990); Berns (1990); Bachman/Palmer (1996); McNamara (1995, 1996, 1997a); Byram (1997); Chapelle/Grabe/Berns (1997); North (1997).

Ein relativ detailliertes und in Teilen empirisch abgesichertes Modell kommunikativer Kompetenz als Basis kommunikativen Sprachgebrauchs innerhalb und außerhalb von Testsituationen findet sich bei Bachman (1990: 84ff.). Eine leicht modifizierte Version wird in Bachman/Palmer (1996: Kap. 4) diskutiert. Bachmans Modell hat sowohl die theoretische Diskussion als auch die empirische Forschung im Bereich des Sprachtestens maßgeblich beeinflußt. Auch eine Reihe von Testentwicklungsprojekten bezieht sich explizit auf das Modell – so z.B. das Projekt "TOEFL 2000" bei der Definition des Konstrukts "communicative language proficiency" (vgl. Chapelle-/Grabe/Berns 1997: 2f.).

Bachman/Palmer (1996) gehen davon aus, daß Sprachfähigkeit ("language ability") als Teil eines interaktiven Modells des Sprachgebrauchs zu explizieren ist. Ihr Modell besteht aus zwei in Wechselwirkung stehenden Hauptkomponenten: 1. Merkmale des Sprachgebrauchs in und außerhalb von Testsituationen; 2. Merkmale der Sprachbenutzer. Unter Sprachgebrauch verstehen die Autoren die „dynamische und interaktive Aushandlung von intendierten Bedeutungen zwischen zwei oder mehr Individuen in einer spezifischen Situation" (S. 60f.). Im Hinblick auf die Entwicklung und den Gebrauch von Sprachtests sehen Bachman/Palmer insbesondere folgende Merkmale der Sprachbenutzer/Testkandidaten als zentral an: 1. Persönlichkeitsvariablen wie Alter, Geschlecht und Muttersprache; 2. Hintergrundwissen ("topical knowledge"); 3. affektive Schemata; 4. Sprachfähigkeit ("language ability"). Kennzeichnend für das Modell ist damit u.a., daß explizit Merkmale der Sprachverwendungssituation und der Testaufgaben sowie persönliche Merkmale des Sprachbenutzers Berücksichtigung finden.

Bei der Sprachfähigkeit differenzieren die Autoren zwischen den folgenden beiden Komponenten: 1. sprachliches Wissen ("language knowledge"; bei Bachman 1990 als "language competence" bezeichnet) und 2. strategische Kompetenz ("strategic competence"). Beide Komponenten werden wiederum – unter Bezug insbesondere auf die grundlegenden Arbeiten von Canale/Swain (1980) und Canale (1983) – in eine Reihe von Teilkomponenten unterteilt. Die Teilkomponenten von „Sprachwissen" sind – geringfügig modifiziert – in Tabelle 1 aufgeführt.

Tabelle 1: Komponenten des Sprachwissens nach Bachman/Palmer (1996: 68)

1. strukturelles Wissen ("organizational knowledge")
(wie Äußerungen und Texte organisiert sind)
1.1 grammatikalisches Wissen
(wie einzelne Äußerungen organisiert sind)
- Lexik
- Syntax
- Phonematik/Graphematik

1.2 textuelles Wissen
(wie Äußerungen in Form von Texten organisiert sind)
- Kohäsion
- Rhetorik

2. pragmatisches Wissen
(wie Äußerungen oder Texte sich auf die kommunikativen Ziele der Sprachbenutzer und den Gebrauchskontext beziehen)
2.1 funktionales Wissen (= "illocutionary competence" bei Bachman 1990)
(wie Äußerungen oder Texte sich auf die kommunikativen Ziele der Sprachbenutzer beziehen)
- Darstellungs- und Ausdrucksfunktion ("ideational function")
- manipulative Funktion: a) instrumentell; b) regulativ; c) interpersonell
- heuristische Funktion
- imaginative Funktion

2.2 soziolinguistisches Wissen
(wie Äußerungen oder Texte sich auf den Gebrauchskontext beziehen)
- Dialekte/Varietäten
- Register
- Idiomatik
- kulturelle Referenzen und Sprechfiguren

Während es sich beim Sprachwissen um spezifische, im Gedächtnis gespeicherte **sprachliche** Wissensbestände handelt, bezeichnet **strategische Kompetenz** bei Bachman und Palmer die **generelle metakognitive Fähigkeit** zum Einsatz sprachlicher und nichtsprachlicher Wissensbestände und Fähigkeiten in einer gegebenen Situation. Strategische Kompetenz besteht aus folgenden drei Hauptkomponenten: 1. **Entscheidung über** die zu verfolgenden **Ziele** ("goal setting"); 2. **Bewertung** ("assessment") der notwendigen Ressourcen sowie der Korrektheit und Angemessenheit von Äußerungen; 3. **Planung** im Sinne einer Formulierung und Auswahl von Plänen und Ressourcen zur Lösung einer Aufgabe.[17]

Strategische Kompetenz trägt insbesondere der Tatsache Rechnung, daß Sprachbenutzer in unterschiedlichem Maße in der Lage sind, ihr Sprachwissen in der Kommunikation zu aktualisieren. Daneben können aber auch bestimmte Fähigkeiten, wie z.B. die Fähigkeit zur Bildung von Inferenzen bei der Bearbeitung eines Lesetests, als direkter Indikator strategischer Kompetenz angesehen werden (vgl. Bachman 1990: 105).

Für Bachman (1990) bzw. Bachman/Palmer (1996) ist die strategische Kompetenz von zentraler Bedeutung im Hinblick auf den Sprachgebrauch innerhalb und außerhalb von Testsituationen. Dies ist ein wichtiger Unterschied zu

[17] Bachman (1990: 100) nennt folgende drei Komponenten strategischer Kompetenz: "assessment", "planning" und "execution". Bei Bachman/Palmer (1996) ist dagegen die **Ausführung** von Plänen selbst nicht länger eine Komponente strategischer Kompetenz (die Ausführungen der Autoren sind allerdings nicht sehr klar in dieser Frage). So heißt es z.B.: "Planning involves the formulation of one or more plans for implementation as a response to the task. Plans are implemented through the performance of language use tasks, involving interpreting and producing utterances or sentences in discourse" (Bachman/Palmer 1996: 79). Skehan (1998: 166) interpretiert dagegen Bachman/Palmer (1996) fälschlicherweise dahingehend, daß wie bei Bachman (1990) die Ausführung eine Teilkomponente strategischer Kompetenz ist. Die weitergehende Frage, ob die Ausführung als Teilkomponente strategischer Kompetenz oder nicht eher als Teil einer on-line-Verarbeitungskomponente im Sinne des dem TOEFL 2000 zugrunde liegenden „Arbeitsmodell kommunikativen Sprachgebrauchs im akademischen Kontext" (vgl. Chapelle/Grabe/Berns 1997) zu betrachten sei, beantwortet Bachman selbst wie folgt: "Off the top of my head, at the moment, I would not consider execution to be part of planning, and probably not part of strategic competence. I find the TOEFL 2000 group's discussion of it in terms of on-line processing reasonably persuasive for now" (persönliche Mitteilung, 3.11.99).

früheren Modellen kommunikativer Sprachkompetenz/Sprachfähigkeit, in denen der strategischen Kompetenz in erster Linie eine kompensatorische Funktion im Fall von Kommunikationsproblemen zugebilligt wurde (so z.B. in dem bekannten Modell von Canale/Swain 1980).

Bachman/Palmer (1996: 62) weisen explizit darauf hin, daß es sich bei ihrem Modell nicht um ein Modell der Sprachverarbeitung, sondern um ein konzeptuelles Schema zur Testkonstruktion und Testinterpretation handelt. Genau an diesem Punkt setzt die Kritik von Skehan (1998) an, der im Hinblick auf das Modell feststellt:

> "Above all, the difficulty is that the account lacks a rationale grounded in psycholinguistic mechanisms and processes (and research findings) which can enable such a model to move beyond 'checklist' status and instead make functional statements about the nature of performance and the way it is grounded in competence." (S. 164)

Skehan selbst favorisiert einen "processing approach", in dem die Wechselwirkung zwischen der Verarbeitungskompetenz der Testkandidaten, den psycholinguistischen Merkmalen der Testaufgaben und dem Verarbeitungskontext der Aufgaben von entscheidender Bedeutung ist. Den Ansatz von Bachman rechnet er hingegen zum sog. "sampling abilities approach" und grenzt beide Ansätze folgendermaßen voneinander ab:

> "An approach based on sampling abilities is likely to use some model of underlying competences, for example ... a model of communicative competence (Canale and Swain 1980). Consequently it will probably regard processing and contexts as things to be handled 'by extension' once the underlying pattern of abilities has been measured. In contrast, a processing approach will regard the capacity to handle real language use as the dominant factor, with abilities playing a subservient, servicing role. The emphasis in testing is then likely to be on establishing a sampling frame, not for abilities, but for the range of performance conditions which operate (for example mode of language use, opportunity to prepare, or degree of time pressure) so that generalizations can be made, in a principled way, to a range of processing conditions." (Skehan 1998: 155)

Die Kritik Skehans an der ungenügenden Berücksichtigung der psycholinguistischen Verarbeitungsdimension bei Bachman (1990), Bachman/ Palmer (1996) sowie anderen Autoren ist sicherlich gerechtfertigt. Hier stellt das

dem TOEFL 2000 zugrundeliegende „Arbeitsmodell kommunikativen Sprachgebrauchs im akademischen Kontext" (vgl. Chapelle/Grabe/Berns 1997: 5) sicherlich einen eindeutigen Fortschritt dar, da im Vergleich zum Bachman-Modell die internen mentalen Verarbeitungsoperationen unter Einschluß der Komponenten „on-line-Verarbeitung" und „verbales Arbeitsgedächtnis" weit stärkere Berücksichtigung finden. Trotz dieser Kritik ist das Modell von Bachman bzw. Bachman/Palmer insbesondere im Hinblick auf die Praxis der Testentwicklung und Testanalyse von beträchtlichem Wert. So werden bei Bachman/Palmer (1996: 253ff.) u.a. eine Reihe von illustrativen Testentwicklungsprojekten mehr oder minder detailliert auf der Basis des Modells spezifiziert. Diese Spezifikationen haben sicherlich ein erhebliches Anregungspotential sowohl im Hinblick auf künftige Testentwicklungsprojekte als auch im Hinblick auf eine kritische Analyse existierender Tests.

9. Eine abschließende Bemerkung zum TESTDAF

Gerade für bedeutende internationale Tests mit einer Vielzahl von Teilnehmern aus unterschiedlichen Ländern ist kennzeichnend, daß die Testaufgaben zwar einerseits möglichst authentisch und valide im Hinblick auf sehr unterschiedliche Adressaten sein sollen, andererseits zugleich jedoch praktikabel und ökonomisch. Dies stellt die Testkonstrukteure vor ein Dilemma. Die Forderung nach Praktikabilität und Ökonomie spricht für geschlossene, objektiv auswertbare Aufgabenformen. Die Forderung nach Authentizität und Validität spricht dagegen für kommunikative Performanztests z.B. in Form von nur wenig gelenkten face-to-face-Interviews. Wie wir jedoch aus der Testtheorie wissen, ist ohne ein gewisses Maß an Objektivität und Reliabilität keine hinreichende Validität und Vergleichbarkeit der Testresultate möglich. Dies bedeutet, daß der entsprechende kommunikative Performanztest hinreichend objektiv und reliabel sein muß. Eine hinreichende Objektivität und Reliabilität läßt sich jedoch im Fall kommunikativer Performanztests nur mit einem relativ hohen Aufwand gewährleisten. Gegen einen hohen Aufwand spricht wiederum die Forderung nach Praktikabilität und Ökonomie.

Vor dem beschriebenen Dilemma stehen auch die Entwickler des TESTDAF. Ich möchte dies anhand des Testteils „mündlicher Ausdruck" verdeutlichen. Wie bereits erwähnt, wird der mündliche Ausdruck im TESTDAF anhand eines "simulated oral proficiency interview" – kurz SOPI – gemessen. Beim SOPI werden die Stimuli per Tonträger präsentiert und die Antworten der Testpersonen auf Tonträger aufgenommen. Es handelt sich um eine international eingesetzte Testform, die zudem in einer Reihe von Studien empirisch untersucht worden ist.

Dem simulierten Interview fehlen sicherlich wichtige Merkmale genuiner Kommunikation. So kann z.B. die Fähigkeit zum manipulativen Sprachgebrauch im Sinne von Bachman/Palmer (1996) sicherlich nur unzureichend mit dem SOPI gemessen werden. Dies schränkt zumindest potentiell dessen Validität ein. Durch die Standardisierung des Inputs und durch die zentrale Korrektur der Tonträger ist jedoch die Objektivität und Reliabilität des Verfahrens und damit – zumindest potentiell – auch dessen Validität höher als die eines traditionellen mündlichen Interviews. So müßte im Fall eines face-to-face Interviews u.a. auch gewährleistet sein, daß weltweit eine genügende Zahl an kompetenten Interviewern zur Verfügung steht. Außerdem ist das SOPI ökonomischer als ein face-to-face-Interview, da es mit mehreren Kandidaten zugleich durchgeführt werden kann (vgl. zu diesen Argumenten auch Shohamy 1994; Kuo/Jiang 1997). Zudem hat man empirisch mehrfach nachweisen können, daß zumindest im unteren und mittleren Leistungsbereich eine relativ hohe Übereinstimmung von Ergebnissen im SOPI und in entsprechenden nicht-simulierten mündlichen Interviews besteht.[11] Diese Argumente und Befunde sprechen m.E. für das gewählte SOPI und entkräften zumindest einen Teil der geäußerten Kritik. Sie entheben die Entwickler des TESTDAF natürlich nicht davon, das gewählte Verfahren in späteren Projektphasen im Hinblick auf die speziellen Adressaten weiter empirisch zu validieren.

Literaturhinweise

Amor, Stuart: Authenticity in the language classroom. In: *Der Fremdsprachliche Unterricht Englisch* 41.5(1999), 4-10.

Bachman, Lyle F.: *Fundamental considerations in language testing.* Oxford: Oxford University Press 1990.

Bachman, Lyle F./Eignor, Daniel R.: Recent advances in quantitative test analysis. In: Clapham, Caroline/Corson, David (Hrsg.): *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer 1997, 227-242.

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

Bailey, Kathleen M.: *Washback in language testing.* Princeton, NJ: Educational Testing Service 1999. (TOEFL Monograph Series MS-15)

Baker, Rosemary: *Classical test theory and item response theory in test analysis.* Special Report No 2: Language Testing Update. Lancaster University: Centre for Research in Language Education 1997.

[11] Vgl. z.B. Stansfield/Kenyon (1992; 1996) und Kenyon/Tschirner (1999). Vgl. allerdings auch z.B. Shohamy (1994), O'Loughlin (1997) und Koike (1998), die anhand empirischer Analysen zu dem Schluß kommen, daß OPI und SOPI u.a. wegen der im SOPI fehlenden sozial.diskursiven Interaktivität keineswegs als äquivalent anzusehen sind.

Berns, Margie: *Contexts of competence: Social and cultural considerations in communicative language teaching.* New York: Plenum 1990.

Brindley, Geoff: Outcomes-based assessment and reporting in language learning programmes: A review of the issues. In: *Language Testing* 15.1(1998), 45-85.

Brown, James D.: *Testing in language programs.* Englewood Cliffs, NJ: Prentice-Hall 1996.

Byram, Michael: *Teaching and assessing intercultural communicative competence.* Clevedon, England: Multilingual Matters 1997.

Canale, Michael: On some dimensions of language proficiency. In: Oller, John W. Jr. (Hrsg.): *Issues in language testing research.* Rowley, Mass.: Newbury House 1983, 333-342.

Canale, Michael/Swain, Merrill: Theoretical bases of communicative approaches to second language teaching and testing. In: *Applied Linguistics* 1.1(1980), 1-47.

Chapelle, Carol A.: Validity in language assessment. In: *Annual Review of Applied Linguistics* 19(1999), 254-272.

Chapelle, Carol/Grabe, William/Berns, Margie: *Communicative language proficiency: Definition and implications for TOEFL 2000.* Princeton, NJ: Educational Testing Service 1997. (TOEFL Monograph Series MS-10)

Chomsky, Noam: *Aspects of the theory of syntax.* Cambridge, MA: MIT Press 1965.

Cronbach, Lee J.: *Essentials of psychological testing.* New York: Harper & Row 1984. (4th edition)

Davies, Alan: Communicative competence as language use. In: *Applied Linguistics* 10.2(1989), 157-170.

Decoo, Wilfried/Colpaert, Josef: Competence and performance in terms of content validity in productive language testing. In: Gardenghi, Monica/O'Connell, Mary (Hrsg.): *Prüfen, Testen, Bewerten im modernen Fremdsprachenunterricht.* Frankfurt am Main: Lang 1997, 25-36.

Douglas, Dan: Language for specific purposes testing. In: Clapham, Caroline/Corson, David (Hrsg.): *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer 1997, 111-119.

Doyé, Peter: Authenticity in foreign language testing. In: *Zielsprache Englisch* 23.2(1993), 1-6.

Edelhoff, Christoph: Authentizität im Fremdsprachenunterricht. In: Edelhoff, Christoph (Hrsg.): *Authentische Texte im Deutschunterricht: Einführung und Unterrichtsmodelle.* München: Hueber 1985, 7-30.

Fischer, Gerhard: *Einführung in die Theorie psychologischer Tests.* Bern: Huber 1974.

Fischer, Gerhard/Molenaar, Ivo W. (Hrsg.): *Rasch models: Foundations, recent developments, and applications.* New York: Springer 1995.

Glaboniat, Manuela: *Kommunikatives Testen im Bereich Deutsch als Fremdsprache: Eine Untersuchung am Beispiel des Österreichischen Sprachdiploms Deutsch.* Innsbruck & Wien: Studien-Verlag 1998.

Grotjahn, Rüdiger: Test validation and cognitive psychology: some methodological considerations. In: *Language Testing* 3.2(1986), 159-185.

Grotjahn, Rüdiger: Psychologie cognitive et évaluation en langue étrangère. In: Association de didactique du français langue étrangère (ASDIFLE) (Hrsg.): *Les cahiers de l'ASDIFLE No 5: Certifications linguistiques en Europe: Problématique, instrumentations, méthodologies. Actes des 11e et 12e Rencontres.* Paris: ASDIFLE 1994, 166-179. (= 1994a)

Grotjahn, Rüdiger: Welche Sprachtests für welchen Zweck? In: Wolff, Armin/Gügold, Barbara (Hrsg.): *Deutsch als Fremdsprache ohne Mauern.* Regensburg: Fachverband Deutsch als Fremdsprache 1994, 31-49. (= 1994b)

Grotjahn, Rüdiger: Strategiewissen und Strategiegebrauch. Das Informationsverarbeitungsparadigma als Metatheorie der L2-Strategieforschung. In: Rampillon, Ute/Zimmermann, Günther (Hrsg.): *Strategien und Techniken beim Erwerb fremder Sprachen.* Ismaning: Huber 1997, 33-76.

Grotjahn, Rüdiger: Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TESTDAF. In: Bolton, Sibylle (Hrsg.): *TESTDAF: Beiträge aus einem Expertenkolloquium.* Köln: VUB-Gilde 2000 (im Druck).

Grotjahn, Rüdiger/Klein-Braley, Christine: Testen. In: Jung, Udo O. H. (Hrsg.): *Praktische Handreichung für Fremdsprachenlehrer.* 2., verb. und erw. Aufl. Frankfurt am Main: Lang 1998, 294-301.

Hamp-Lyons, Liz: Ethics in language testing. In: Clapham, Caroline/Corson, David (Hrsg.): *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer 1997, 323-333.

Hughes, Arthur: *Testing for language teachers.* Cambridge: Cambridge University Press 1989.

Hymes, Dell: Competence and performance in linguistic theory. In: Huxley, Renira/Ingram, Elisabeth (Hrsg.): *Language acquistion: Models and methods.* London & New York: Academic Press 1971, 3-28.

Hymes, Dell: On communicative competence. In: Pride, John B./Holmes, Janet (Hrsg.): *Sociolinguistics.* Harmondsworth: Penguin 1972, 269-293.

Ingenkamp, Karlheinz: *Lehrbuch der Pädagogischen Diagnostik* (Studienausgabe). Weinheim & Basel: Beltz 1985.

Kenyon, Dorry/Tschirner, Erwin: *The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels.* Ms. 1999

Kieweg, Werner: Allgemeine Gütekriterien für Lernzielkontrollen. In: *Der Fremdsprachliche Unterricht Englisch* 33.1(1999), 4-11.

Klauer, Karl J.: *Kriteriumsorientierte Tests.* Göttingen: Hogrefe 1987.

Klein-Braley, Christine: Ask a stupid question ...: Testing language proficiency in the context of research studies. In: de Bot, Kees/Ginsberg, Ralph/Kramsch, Claire (Hrsg.): *Foreign language research in cross-cultural perspective.* Amsterdam & Philadelphia: Benjamins 1991, 73-94.

Koike, Dale A.: What happens when there's no one to talk to? Spanish foreign language discourse in simulated oral proficiency interviews. In: Young, Richard/He, Agnes W. (Hrsg.): *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam: Benjamins 1998, 69-98.

Krauth, Joachim: *Testkonstruktion und Testtheorie.* Weinheim: Psychologie Verlags Union 1995.

Kunnan, Antony J.: Approaches to validation in language assessment. In: Kunnan, Antony J. (Hrsg.): *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach.* Mahwah, NJ: Erlbaum 1998, 1-16.

Kunnan, Antony J.: Recent developments in language testing. In: *Annual Review of Applied Linguistics* 19(1999), 235-253. (=1999a)

Kunnan, Antony J. (Hrsg.): *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida.* Cambridge: Cambridge University Press 1999. (=1999b)

Kuo, Jane/Jiang, Xixiang: Assessing the assessments: the OPI and the SOPI. In: *Foreign Language Annals* 30.4(1997), 503-512.

Lee, Wonkyung: *Prüfungen zum Nachweis deutscher Sprachkenntnisse bei ausländischen Studienbewerbern/Studienbewerberinnen (PNdS/DSH): Ihre Praxis und ihr Prüfprofil.* Regensburg: Fachverband Deutsch als Fremdsprache 1998.

Lewkowicz, Jo A.: Autheticity for whom? Does authenticity really matter? In: Huhta, Ari/Kohonen, Viljo/Kurki-Suonio, Lisa/Luoma, Sari (Hrsg.): *Current developments and alternatives in language assessment: Proceedings of LTRC 96.* Jyväskylä: University of Jyväskylä 1997, 165-184. (= 1997a)

Lewkowicz, Jo A.: *Investigating authenticity in language testing.* Unpublished doctoral dissertation. University of Lancaster 1997. (= 1997b)

Lienert, Gustav A./Raatz, Ulrich: *Testaufbau und Testanalyse.* Weinheim: Beltz, Psychologie Verlags Union 1994.

McNamara, Tim F.: Modelling performance: Opening Pandora's box. In: *Applied Linguistics* 16(1995), 159-179.

McNamara, Tim F.: *Measuring second language performance.* London: Longman 1996.

McNamara, Tim F.: 'Interaction' in second language performance assessment: Whose performance. In: *Applied Linguistics* 18(1997), 446-466. (= 1997a)

McNamara, Tim F.: Performance testing. In: Clapham, Caroline/Corson, David (Hrsg.): *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer 1997, 131-139. (= 1997b)

McNamara, Tim F.: Policy and social considerations in language assessment. In: *Annual Review of Applied Linguistics* 18(1998), 304-319.

Messick, Samuel: Validity. In: Linn, Robert L. (Hrsg.): *Educational measurement (3rd ed.).* New York: American Council on Education/Macmillan Publishing Company 1989, 1-103.

Messick, Samuel: The interplay of evidence and consequences in the validation of performance assessment. In: *Educational Researcher* 23.2(1994), 13-23.

Messick, Samuel: Validity and washback in language testing. In: *Language Testing* 13(1996), 241-256.

Moss, Pamela A.: Shifting conceptions of validity in educational measurement: Implications for performance assessment. In: *Review of Educational Research* 62(1992), 229-258.

Moss, Pamela A.: Can there be validity without reliability? In: *Educational Researcher* 23.2(1994), 5-12.

Norris, John M./Brown, James D./Hudson, Thom/Yoshioka, Jim: *Designing second language performance assessments.* University of Hawai'i: Second Language Teaching & Curriculum Center 1998. (Technical Report #18)

North, Brian: Perspectives on language proficiency and aspects of competence. In: *Language Teaching* 30.2(1997), 93-100.

Oller, John W. Jr.: *Language tests at school.* London: Longman 1979.

O'Loughlin, K.: *The equivalence of two versions of an oral proficiency test.* Unpublished PhD thesis. University of Melbourne 1997.

Peirce, Bonny N.: Demystifying the TOEFL reading test. In: *TESOL Quarterly* 26.4(1992), 665-691.

Pollitt, Alastair: Rasch measurement in latent trait models. In: Clapham, Caroline/Corson, David (Hrsg.): *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer 1997, 243-253.

Rost, Jürgen: *Lehrbuch Testtheorie, Testkonstruktion.* Bern: Huber 1996.

Schwarzer, Christine/Schwarzer, Ralf: Informelle Tests. In: Klauer, Karl J. (Hrsg.): *Handbuch der Pädagogischen Diagnostik. Studienausgabe.* Düsseldorf: Schwann 1982, 317-330.

Shohamy, Elana: The validity of direct versus semi-direct oral tests. In: *Language Testing* 11.2(1994), 99-123.

Sigott, Günther: Language test validity: An overview and appraisal. In: *Arbeiten aus Anglistik und Amerikanistik* 19.2(1994), 287-294.

Skehan, Peter: *A cognitive approach to language learning.* Oxford: Oxford University Press 1998.

Spolsky, Bernard: The limits of authenticity in language testing. In: *Language Testing* 2.1(1985), 31-40.

Spolsky, Bernard: Communicative competence, language proficiency, and beyond. In: *Applied Linguistics* 10.2(1989), 138-156.

Stansfield, Charles W./Kenyon, Dorry M.: Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. In: *System* 20.3(1992), 347-364.

Stansfield, Charles W./Kenyon, Dorry M.: *Simulated Oral Proficiency Interviews: An update (http://www.cal.org/ericcll/digest/stansf02.html).* May 1996.

Stumpf, Heinrich: Klassische Testtheorie. In: Erdfelder, Edgar/Mausfeld, Rainer/Meiser, Thorsten/Rudinger, Georg (Hrsg.): *Handbuch Quantitative Methoden.* Weinheim: Psychologie Verlags Union 1996, 411-430.

Taylor, David A.: The meaning and use of the term 'competence' in linguistics and applied linguistics. In: *Applied Linguistics* 9.2(1988), 148-168.

Trim, J. L. M./North, Brian (Hrsg.): *Transparency and coherence in language learning in Europe: Objectives, evaluation, certification.* Strasbourg: Council of Europe 1992.

van der Linden, Wim J./Hambleton, Ronald K. (Hrsg.): *Handbook of modern item response theory.* New York: Springer 1997.

Wall, Dianne: Impact and washback in language testing. In: Clapham, Caroline/ Corson, David (Hrsg.): *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer 1997, 291-302.

Widdowson, Henry G.: Knowledge of language and ability for use. In: *Applied Linguistics* 10.2(1989), 128-137.

Young, Richard/He, Agnes W. (Hrsg.): *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam: Benjamins 1998.