MODUL 1

MODELLE DER SPRACHKOMPETENZ

MODELLE DER SPRACHKOMPETENZ

Die angewendeten Testverfahren reflektieren im Allgemeinen die jeweils gültige Auffassung von Sprache und Sprachgebrauch. Was geprüft werden soll und die dafür verwendeten Aufgaben oder Items sind von den derzeit gültigen Vorstellungen darüber beeinflusst, was Sprachkompetenz ist und was genau wir tun, wenn wir Sprache in Alltagssituationen gebrauchen.

1. Der psychometrisch-strukturalistische Ansatz

Der vorherrschende Ansatz bei der Erstellung von Fremdsprachentests hielt sich bis zum Ende der 70er Jahre an die Prinzipien, die von Wissenschaftlern wie Robert Lado Ende der 60er Jahre formuliert wurden. So wie der Strukturalismus in seinen verschiedenen Ausprägungen während dieser Zeit die Basis für die Entwicklung von Unterrichtsmaterialien bildete, galt er auch als Grundlage für die Curricula und für die Testentwicklung. Die Grundprinzipien für diese Art von Fremdsprachentests, die auf der psychologischen Theorie des Behaviorismus und der linguistischen Theorie des Strukturalismus beruhten, zeigen sich deutlich in den Arbeiten von Lado (1961), Valette (1967), Harris (1969) und Heaton (1975).

Testaufgaben und -items aus dieser Zeit, die oft als die psychometrisch-strukturalistische Ära (Spolsky, 1975) bezeichnet wird, sind besonders dadurch gekennzeichnet, dass sie die Objektivität der Bewertung betonen. Diese wurde durch die Verwendung sorgfältig erstellter (dekontextualisierter) Einzelaufgaben im Multiple-Choice-Format erreicht. Selbst produktive Fertigkeiten wie das Schreiben oder die Aussprache wurden indirekt oder rezeptiv im Multiple-Choice-Format überprüft. Die erstellten Tests folgten implizit einer hierarchischen Sicht von Sprachkompetenz, die im Einklang mit der strukturalistischen Auffassung von Sprache stand, d. h. der Auffassung, dass sich Sprache aufbaut – vom Phonem zum Morphem, zum Wort und zum Satz. Es gab kaum einen Versuch, die Sprachkompetenz explizit zu beschreiben, obwohl Lado ansatzweise in diese Richtung ging, indem er den Spracherwerbsprozess als Internalisierung einer Reihe von Gewohnheiten beschrieb:

> „These habits involve matters of form, meaning and the distribution of layers of structure, namely those of the sentence, clause, phrase, word, morpheme, and phoneme." (Lado, 1961: 22)

a) Kritik am strukturalistischen Ansatz

Obwohl solche Tests mit dekontextualisierten Items in den 60er und 70er Jahren dominierten und sogar heutzutage noch in einigen Ländern der Welt eingesetzt werden, wurden sie auch damals schon kritisch gesehen. Carroll (1961) schrieb, dass eine wesentliche Einschränkung dekontextualisierter Einzelaufgaben darin liegt, dass ein Item jeweils nur ein Element der Sprache erfassen kann und dass dies in den meisten Fällen nicht dem wirklichen Gebrauch von Sprache entspricht. Er schlug vor, Tests einzusetzen, die die kommunikative Wirkung von Äußerungen erfassen und nicht isolierte Elemente der Sprache. Carroll nannte solche Tests integrativ. Die folgende Äußerung verdeutlicht, was er mit integrativen Tests meinte:

> „Since the use of language in ordinary situations calls upon all these aspects (of language), we must further recognize that linguistic performance also involves the individual's capability of mobilizing his linguistic competence and performance abilities in an integrated way, i.e. in understanding, speaking, reading or writing in connected discourse." (Carroll, 1968: 58)

Fragen danach, wer mit wem, warum und in welcher Situation kommuniziert, sowie das Ausmaß, in dem die Performanz auf der zugrunde liegenden Kompetenz beruht, standen für die damalige Testkonstruktion nicht im Mittelpunkt, obwohl man nicht sagen kann, dass sie überhaupt keine Rolle spielten. Die linguistischen Aspekte der Sprachkompetenz waren nur einfacher zu isolieren und somit einfacher zu prüfen. Darüber hinaus bekam die Linguistik durch den Ansatz von Chomsky eine neue Richtung und beschäftigte sich nun mit der Analyse der Sprache und dem idealen Sprecher/Hörer. Die Rolle von Sprache als Mittel der Kommunikation wurde nie gründlich analysiert. Das folgende Zitat von Lado verdeutlicht seine Auffassung, dass das Testen von Sprache als Mittel der Kommunikation ein viel zu kompliziertes Unterfangen sei, um in Angriff genommen zu werden:

> „The situations in which language is the medium of communication are potentially almost infinite. No one, even the most learned, can speak and understand his native language in all the situations in which it can be used ... even if we could pick only valid situations and even if we could be sure that understanding these situations occurred through the language used, we could still have the problem of the great variety of situations which must be sampled. The elements of the language on the other hand are limited, and it is more profitable to sample these elements than the great variety of situations in which the language is used." (Lado, 1961)

Testautoren konnten Items erstellen, die den Kriterien der Leistungsmessung und der Linguistik entsprachen, sie unternahmen jedoch keinen nennenswerten Versuch, einen validen äußeren Kontext herzustellen. Die Begründung für die Erstellung von Items lag darin, dass sie einer Liste von Strukturen entsprachen, nicht dass sie den Gebrauch von Sprache in Realsituationen widerspiegelten. Da jedoch auch die Unterrichtsmethoden dem Strukturalismus folgten, wurden solche Tests als valide angesehen. Ihre Validität kann heute mit dem Argument in Zweifel gezogen werden, dass die damit erfassten sprachlichen Elemente weder angemessen noch authentisch waren und dass der Zusammenhang zwischen „use" (Anwendung, Gebrauch) und „usage" (formales System der Sprache) nicht berücksichtigt wurde (Alderson, 1981).

Trotz dieser Einwände sollte man nicht vergessen, welchen wichtigen Beitrag die psychometrisch-strukturalistische Ära für die Testkonstruktion geleistet hat. Die Betonung statistischer Analysen, der Reliabilität und Validität, der Planung von Testinhalten in Bezug auf die sprachlichen Strukturen sowie die Entwicklung von (dekontextualisierten) Multiple-Choice-Items waren von dauerhaftem Wert.

2. Die pragmatische Wende

Kommunikative Sprachtests entwickelten sich aus der Abkehr des Sprachunterrichts, der Didaktik und der Sprachlerntheorie vom vorwiegend strukturalistischen Ansatz hin zum Gebrauch der Sprache in Realsituationen.

Dieser neue Ansatz vollzog sich zunächst in der Linguistik und er wurde durch Entwicklungen in benachbarten Disziplinen, wie der Soziolinguistik, fortgeführt und modifiziert. Hymes (1967) entwickelte den Begriff des Sprechakts, ein Begriff, der deutlich machen soll, dass sprachliches Handeln von Regeln des Gebrauchs bestimmt wird. Er betonte, dass die verschiedenen Sprechakte durch unterschiedliche Regeln des Gebrauchs bestimmt werden und dass man ihre Struktur dadurch definieren kann, dass man sie in einzelne am Sprechakt beteiligte Faktoren untergliedert, wie Kommunikationsteilnehmer, Kontext und Situation, Kommunikationsziel, Thema, Kanal usw. Hymes führte den Begriff der kommunikativen Kompetenz ein und hob dadurch hervor, dass eine angemessene sprachliche Kompetenz mehr beinhaltet als nur sprachliches Wissen:

> „There are rules of use without which the rules of grammar would be useless. Just as rules of syntax can control aspects of phonology, just as rules of semantics perhaps control aspects of syntax, so rules of speech acts enter as a controlling factor for linguistic form as a whole." (Hymes, 1972: 278)

Für den Sprachunterricht wurden diese Ansichten am ausführlichsten von Munby (1978) ausgearbeitet. Sein Ansatz beruht auf der Prämisse, dass die zu unterrichtende Sprache so eng wie möglich mit den unmittelbaren und zukünftigen Bedürfnissen der Lerner in Verbindung gebracht wird, dass die Lerner auf eine authentische Kommunikation vorbereitet werden und dass die unterrichtete Sprache einen hohen Nutzwert erhält (Wilkins, 1976). Diesem Ansatz folgte auch van Ek in seiner Lernzielbeschreibung (*engl. specification*) für den Threshold Level, der 1975 vom Europarat veröffentlicht wurde. Eine von van Ek und Trim revidierte Version erschien unter dem Titel „Threshold Level 1990".

3. Der Beitrag des Threshold Level

Der Threshold Level, der dem kommunikativen Ansatz verpflichtet war, hatte einen weit reichenden und dauerhaften Einfluss sowohl auf den Unterricht als auch auf die Testentwicklung. In der Einleitung zur Ausgabe von 1980 heißt es, dass sich für den Sprachunterricht ein funktionaler Ansatz empfiehlt, um den Unterricht von einer „strukturalistisch dominierten wissenschaftlichen Sterilität zu befreien und zu einem lebendigen Medium zu machen, das der größeren Mobilität von Menschen und Gedanken dient". Bei diesem Ansatz liegt die Betonung auf dem Gebrauch von Sprache als Kommunikationsmittel und den alltäglichen kommunikativen Bedürfnissen eines Erwachsenen, der sich in einem fremden Land aufhält.

Der Threshold Level ist kein Kurs oder Curriculum und auch keine umfassende Liste der grammatischen und lexikalischen Elemente, die ein Lerner auf einer bestimmten Stufe beherrschen muss, sondern eine Auflistung von Zielen, ein Versuch „zu spezifizieren, wie ein Lerner die Sprache gebrauchen sollte, um sich unabhängig in einem Land zu bewegen, in dem diese Sprache das Kommunikationsmittel im Alltag ist". Dies bedeutet, dass der Lerner nicht nur in die Lage versetzt werden muss, Einkäufe zu machen oder sein Auto reparieren zu lassen, sondern auch Informationen und Meinungen mit anderen Menschen auszutauschen, über Vorlieben und Abneigungen sowie über eigene Erfahrungen zu sprechen. Die Betonung liegt auf Sprache als sozialem Instrument, als einem Mittel, das Menschen dazu befähigt, miteinander zu interagieren. Den Ausgangspunkt bilden Situationen, in denen sich Sprachenlerner üblicherweise in einem fremden Land befinden; das Ziel besteht darin, die Sprache so zu beherrschen, dass ein kommunikativ angemessenes Handeln in diesen Situationen möglich ist.

In den Zielbeschreibungen des Threshold Level sind die Elemente der Sprache nicht nach ihrer grammatischen Struktur klassifiziert, sondern unterteilt in Funktionen und Begriffe, die sich darauf beziehen, was Menschen mittels der Sprache tun, also auf kommunikative Aktivitäten. Die Funktionen werden in sechs Hauptkategorien unterteilt: Informationen geben und erfragen; Einstellungen ausdrücken; Handlungsziele verwirklichen; Sozialkontakte pflegen; den Diskurs strukturieren; Kommunikation verbessern. Jede dieser sechs Kategorien kann weiter unterteilt werden. Zum Beispiel beinhaltet „Informationen geben und erfragen" die Unterkategorien „identifizieren", „erzählen/berichten" und „korrigieren" und es werden jeweils einige beispielhafte Versprachlichungen aufgeführt (z. B. unter „identifizieren" „Er ist der Besitzer des Restaurants.").

Außerdem werden acht allgemeine Begriffe genannt: Existenz, Raum, Zeit, Quantität, Qualität, Eigenschaften, Relationen und Deixis. „Existenz" wird untergliedert in Kategorien wie „Anwesenheit/Abwesenheit", „Verfügbarkeit/Nicht-Verfügbarkeit", und es werden jeweils sprachliche Beispiele aufgeführt. Außerdem gibt es eine umfangreiche Liste von spezifischen Begriffen, deren Versprachlichungen unter 14 Themen aufgeführt werden: Person, Wohnung/Haus, Umgebung, Tägliches Leben, Freizeit, Unterhaltung, Reisen, Beziehungen zu anderen Menschen, Gesundheit und Körperpflege, Erziehung, Einkaufen, Essen und Trinken,

Dienstleistungen, Orte, Sprache und Wetter.

Der Threshold Level 1990 enthält auch einen Überblick über die Grammatik und eine Wortschatzliste. Es wird aber betont, dass „dieser Apparat zur Bildung von Sätzen, die Grammatik und die Lexik, kein Ziel ansich ist, sondern ein Mittel, um die kommunikativen Funktionen sprachlich zu realisieren, und dass nur Letzteres wirklich zählt". Ein letzter Aspekt des Threshold, der nicht unerwähnt bleiben sollte, ist seine Flexibilität. Die Versprachlichungen der Funktionen und Begriffe sind nur Beispiele, die durch andere ersetzt werden können. Die Zielbeschreibungen können, z. B. durch Auswahl einer entsprechenden Lexik, an bestimmte Adressatengruppen angepasst werden.

Auch die Eignung des Threshold Level als Vorlage für Übertragungen in andere europäische Fremdsprachen zeigt seine Flexibilität. Diese Übertragungen sind keine Übersetzungen des englischen Originals, sondern unabhängige Interpretationen des Threshold-Konzepts, die die kulturellen Unterschiede berücksichtigen und die erheblich in ihren Kategorien und Versprachlichungen abweichen können. Bei Erscheinen des Threshold Level 1990 gab es – neben Englisch – zehn Versionen in anderen Sprachen und seitdem wurden noch weitere erstellt. Alle sind über den Europarat in Straßburg oder über seine Ländervertretungen in Europa und darüber hinaus erhältlich.

Der Threshold Level und die Vorstellungen, auf denen er beruht, beeinflussten den Unterricht dahingehend, dass man nicht länger ausschließlich Sprachstrukturen vermittelte, sondern nun von Situationen und handlungsorientiertem Lernen ausging und Materialien verwendete, die entweder authentisch oder wenigstens semi-authentisch waren. Es wurden Aufgaben mit „Informationslücken" verwendet, um bei Lernern ein Bedürfnis zu entwickeln, miteinander zu kommunizieren. Auch in der Testentwicklung führte dies zu mehr authentischen (oder semi-authentischen) Textvorlagen, weg von dekontextualisierten Einzelaufgaben hin zu Aufgaben, die in einen kommunikativen Kontext eingebettet sind.

Die Betonung kommunikativer Aktivitäten bringt die Notwendigkeit mit sich, die Fertigkeiten und Fähigkeiten eines Lerners und ihr Zusammenspiel genau zu analysieren, damit er ein bestimmtes Niveau an kommunikativer Kompetenz in einer Sprache erreichen kann, sei es in der Muttersprache oder in einer später angeeigneten Fremdsprache. Während die Testentwicklung in der psychometrisch-strukturalistischen Ära (Spolsky, 1975) den Dimensionen der sprachlichen Fähigkeit und der kommunikativen Kompetenz wenig Aufmerksamkeit schenkte, konzentrierte sie sich seit Mitte der 70er Jahre auf diesen Bereich.

4. Modelle kommunikativer Kompetenz

Modelle kommunikativer Kompetenz stellen seit den späten 70er Jahren einen wichtigen Schwerpunkt des Unterrichts und der Testentwicklung dar. Im Hinblick auf die Testforschung dieser Zeit nahm Cziko (1982) eine sinnvolle Unterscheidung vor. Er teilt die Forschung in zwei große Kategorien auf, die er beschreibende Modelle und prozessorientierte Modelle kommunikativer Kompetenz nennt. Beschreibende Modelle zeichnen sich dadurch aus, dass sie

- alle Komponenten des Wissens und der Fertigkeiten beschreiben, die man benötigt, um effektiv und angemessen in einer bestimmten Sprache zu kommunizieren.

Prozessorientierte Modelle zeichnen sich dadurch aus, dass sie versuchen aufzuzeigen,

- wie die Komponenten der kommunikativen Kompetenz psychologisch miteinander verbunden sind, um ein Bündel unabhängiger Faktoren zu bilden.

Zu den beschreibenden Modellen gehören die Arbeiten von Canale, Swain und Cummins, während Oller, Palmer, Bachman und andere prozessorientierte Modelle entwickelten.

a) Das Modell von Canale und Swain

Das wohl bekannteste beschreibende Modell der kommunikativen Kompetenz wurde in den 80er Jahren von Canale und Swain (1981; 1983) entwickelt. In diesem Modell umfasst die kommunikative Kompetenz vier Komponenten: die linguistische, die soziolinguistische, die diskursive und die strategische Kompetenz. Die linguistische Kompetenz bezieht sich auf die Beherrschung des formalen Systems der Sprache (Wortschatz, Morphologie, Syntax, Aussprache, Rechtschreibung). Die soziolinguistische Kompetenz trägt zur Fähigkeit des Individuums bei, kontextangemessen zu kommunizieren. Diese Kategorie beschreibt das Ausmaß, in dem sprachliche Äußerungen in jeweils verschiedenen Situationen und in Abhängigkeit von Faktoren, wie Zweck der Interaktion, Status bzw. Rolle der Beteiligten usw., angemessen produziert und verstanden werden. Sie beinhaltet ein Bewusstsein von kulturspezifischen Regeln sozialer Interaktion. Die diskursive Kompetenz bezieht sich auf die Fähigkeit, grammatische Formen und Bedeutungen so zusammenwirken zu lassen, dass sie einen einheitlichen gesprochenen oder geschriebenen Text bilden. Die strategische Kompetenz schließlich bezieht sich auf die Beherrschung von verbalen und nonverbalen Kommunikationsstrategien, mit deren Hilfe ein Abbruch der Kommunikation verhindert bzw. kompensiert wird. Hierzu zählen Strategien wie Wiederholung, Paraphrase oder langsameres Sprechen. Die strategische Kompetenz unterscheidet sich dadurch von den anderen, von Canale und Swain genannten Kompetenzen, dass sie frei mit ihnen interagiert. Dieses Modell der kommunikativen Kompetenz erweiterte in den späten 80er Jahren den Blick der Testentwickler erheblich, da es den Rahmen für eine Beschreibung und somit eine Validierung lieferte, den es vor dieser Zeit nicht gab.

b) Cummins Modell der kommunikativen Kompetenz

Ein anderes beschreibendes Modell der kommunikativen Kompetenz, das die Testkonstruktion und die Interpretation der Testergebnisse beeinflusste, war das Modell von Cummins (1979; 1983). Sein erstes Modell unterschied zwischen schulisch-akademischer Sprachfähigkeit (CALP: cognitive/academic language proficieny) und alltagssprachlichen Fähigkeiten (BICS: basic interpersonal communication skills). Während jeder im Besitz von BICS sein soll, gilt dies nicht für CALP, die eng mit der Lese- und Schreibfertigkeit zusammenhängt. BICS ist gewissermaßen eine Art Minimalkompetenz, während CALP durch Schulbildung erworben wird. Deshalb, so Cummins, benötigten Lerner mit minoritätssprachlichem Hintergrund sehr viel mehr Zeit, um im Englischen ein ihrem Alter gemäßes Niveau in den schulisch-akademischen Fähigkeiten zu erreichen, als dies bei der direkten Kommunikation der Fall ist.

Cummins (1983) entwickelte seinen Ansatz weiter, indem er darauf hinwies, dass Sprachkompetenz entlang zweier Kontinua gesehen werden kann:

Zunächst gibt es ein Kontinuum, das die mögliche kontextabhängige Unterstützung beim Empfangen oder Senden einer Äußerung betrifft. Die Extreme dieses Kontinuums werden als „context-embedded“ (in einen Kontext eingebettete) im Gegensatz zu „context-reduced“ (kontext-reduzierte) Kommunikation bezeichnet. Beide unterscheiden sich durch die Tatsache, dass die Kommunikationspartner bei einer in einen Kontext eingebetteten Kommunikation Bedeutungen aktiv verhandeln können (z. B., indem sie Rückmeldung geben, dass die Äußerung nicht verstanden wurde). Kontextreduzierte Kommunikation hingegen stützt sich vorwiegend (und im Extrem des Kontinuums sogar ausschließlich) auf sprachliche Signale, um Bedeutung zu vermitteln. In einigen Fällen kann dies sogar bedeuten, das Weltwissen auszuschalten, um die Logik der Kommunikation angemessen zu interpretieren (oder zu manipulieren).

Nach Cummins ist die zwischenmenschliche Kommunikation normalerweise in einen Kontext eingebettet, während die kontextreduzierte Kommunikation in Situationen auftritt, in denen sprachliche Präzision äußerst wichtig ist. Natürlich hängt das Ausmaß, in dem die Kommunikation in einen Kontext eingebettet bzw. kontextreduziert ist, stark von den beteiligten Kommunikationspartnern ab. Aber impliziert wird, dass die Anforderungen der Kommunikation umso höher sind, je weniger Kontext vorhanden ist. Diese Auffassung hat Konsequenzen für die Art der Aufgaben, die in einem Test gestellt werden, da deren Lösung in Beziehung zu dem Umfang an Kontext steht, den der Lerner in den Test einbringt. Das bestätigt die Auffassung, dass

Tests nicht jedem gegenüber gleich fair sein können. Dies sollten Testkonstrukteure in ihrer täglichen Arbeit akzeptieren und bei der Auswahl der Textvorlagen und Aufgaben entsprechend bedenken. Das Gesagte unterstützt die weit verbreitete Ansicht, dass Aufgaben ohne einen dem Testteilnehmer bekannten Kontext schwieriger sein könnten.

Das zweite Kontinuum im Modell von Cummins betrifft das Ausmaß der eigenen kognitiven Beteiligung an einer Aufgabe oder Tätigkeit. Cummins definiert kognitive Beteiligung als die Menge an Information, die simultan oder in kurzer Folge verarbeitet werden muss, um eine Tätigkeit auszuführen.

Testaufgaben werden nach Kontexteinbettung und kognitiven Anforderungen kategorisiert. Es ist jedoch nicht einfach, den Begriff „Kontext" präzise zu definieren, da er von Individuum zu Individuum unterschiedlich verstanden wird. Darüber hinaus kann das, was zu Beginn eines Lernabschnitts kognitiv anspruchsvoll ist, zu einem späteren Zeitpunkt nicht mehr anspruchsvoll sein.

Eine der wichtigsten Konsequenzen dieses Modells für die Testentwicklung besteht darin, dass es Testentwickler ermutigt, mehr auf den Lerner als Individuum zu achten. Um die kontextabhängigen sowie die kognitiven Aspekte zu berücksichtigen, muss der Testentwickler die Voraussetzungen der Kandidaten bedenken.

c) Morrow und authentische Kommunikation

Morrow (1979) versucht nur ansatzweise, kommunikative Kompetenz zu definieren, dennoch sind seine Auffassungen für das Testen relevant, da er einige Aspekte authentischer Kommunikation aufzählt, die bei kommunikativen Tests berücksichtigt werden sollten, damit diese valide sind:

1. Kommunikation beruht insofern auf Interaktion, als das, was jemand sagt oder schreibt, unmittelbar davon abhängt, was ihm/ihr gesagt bzw. geschrieben wird.

2. Kommunikation ist unvorhersehbar und die Informationsmenge muss in Echtzeit verarbeitet werden.

3. Kommunikation erfordert einen Kontext, der sowohl die Sprache als auch die Situation betrifft.

4. Kommunikation ist insofern zielorientiert, als eine Person erkennen muss, warum Äußerungen an sie gerichtet werden, und relevante Antworten geben muss, die die gewünschte Wirkung erzielen.

5. Kommunikation erfordert Performanz, d. h. die Fähigkeit, Sprache in Realsituationen gebrauchen zu können.

6. Kommunikation erfordert den Einsatz authentischer Texte.

7. Kommunikation basiert insofern auf Verhalten, als sie ein Ergebnis hat.

i) Nachteile von Morrows Ansatz

In seinem Artikel von 1979 „Communicative language testing: Revolution or evolution?" erklärt Morrow diese Aspekte ausführlicher. Obwohl dies eine nützliche und für die Konstruktion kommunikativer Tests wertvolle Liste ist, wurde sie kritisiert (Alderson, 1981; Weir, 1981; Moller, 1981), da Morrow weder die Begriffe kommunikative Sprachfähigkeit, Sprachkompetenz, Performanztest und Verhaltensergebnis definiert noch hinreichend erklärt, wie die sieben oben aufgeführten Aspekte beim Entwurf kommunikativer Tests angemessen berücksichtigt werden können und wie man sie messen und gewichten sollte. Jedoch hat dieser Ansatz seit den späten 70er Jahren Tests hervorgebracht, die ansprechender und realistischer aussehen, also eine hohe Augenscheinvalidität haben.

d) Bachman und die kommunikative Kompetenz

In den späten 80er Jahren gab es im Sprachtestbereich eine Reihe von neuen Entwicklungen. Bachman (1990) veröffentlichte sein erstes Modell der kommunikativen Sprachkompetenz (CLA: communicative language ability), das eindeutig von den Arbeiten von Canale und Swain beeinflusst war. Er geht davon aus, dass es sich zusammensetzt aus Wissen bzw. Kompetenz sowie der Fähigkeit, diese Kompetenz in angemessene Sprachverwendung umzusetzen. Kommunikative Kompetenz teilt sich nach Bachman in drei Bereiche auf: in Sprachkompetenz, strategische Kompetenz sowie psycho-physiologische Mechanismen. Diese interagieren mit dem Kontext der Sprachverwendung und den Wissensstrukturen des Sprachverwenders.

Kommunikative Kompetenz lässt sich in strukturelles Wissen und pragmatisches Wissen unterteilen. Strukturelles Wissen ist weiter unterteilt in grammatisches Wissen (Lexik, Morphologie, Syntax und Phonologie/Orthographie) und textuelles Wissen (Kohäsion und rhetorische Organisation von gesprochenen und geschriebenen Texten). Die pragmatische Kompetenz betrifft den Zusammenhang zwischen den in der Kommunikation gegebenen sprachlichen Signalen und dem Sprachverwender sowie dem Kontext der Kommunikation. Sie unterteilt sich in illokutive und soziolinguistische Kompetenz. Die illokutive Kompetenz betrifft die Fähigkeit des Sprachverwenders, Sprachfunktionen umzusetzen: kognitive/affektive (um Ideen, Wissen, Gefühle auszudrücken), manipulative (um die ihn umgebende Welt zu beeinflussen), heuristische (um sein Weltwissen zu erweitern) und imaginative (um mit der Sprache zu spielen, Witze zu machen, Gedichte zu verfassen usw.). Soziolinguistische Kompetenz umfasst den angemessenen Gebrauch der Sprache in einem bestimmten Kontext. Dies bedeutet, über ein Gespür für dialektale Unterschiede und Sprachvarietäten sowie für Register und Natürlichkeit (wie sich ein Muttersprachler ausdrücken würde) zu verfügen und die Fähigkeit zu besitzen, kulturelle Bezüge und Redewendungen zu verstehen.

In diesem Modell von Bachman interagieren Sprachkompetenz, strategische Kompetenz (die in der Fähigkeit besteht, in der Interaktion die angemessene Sprachverwendung zu bewerten, zu planen und auszuführen) und psycho-physiologische Kompetenz (welche die anatomischen Voraussetzungen einschließt).

In den oben skizzierten Modellen der Sprachkompetenz ist ein Punkt von besonderem Interesse, nämlich die unterschiedliche Definition der strategischen Kompetenz. Es zeigt sich deutlich, dass alle, die darüber schreiben, sie für eine wichtige Kompetenz halten, aber sie ist vermutlich der am schwierigsten zu beschreibende Kompetenzbereich.

Der Begriff „strategische Kompetenz“ wurde zum Teil sehr unterschiedlich gebraucht. Wie oben erwähnt, sehen ihn Canale und Swain als die Fähigkeit, sowohl verbale als auch nonverbale Mittel einzusetzen, um Zusammenbrüche in der Kommunikation zu beheben. Wenn ein Sprecher z. B. bemerkt, dass er nicht verstanden wurde, kann er dies dadurch kompensieren, dass er langsamer oder klarer spricht oder er kann den gleichen Inhalt mit anderen Worten ausdrücken. Eine ähnliche Auffassung findet sich im Threshold Level in dem Abschnitt „Redeorganisation und Verständnissicherung“, der beschreibt, wie der Lerner Strategien erlernen muss, um angemessen mit unvorhergesehenen Anforderungen sowie mit Gedächtnislücken umzugehen. Für Bachman hingegen ist die strategische Kompetenz weniger ein Mittel zum Ausgleich von Mängeln; er sieht sie positiver als die Fähigkeit zur Bewertung und Planung, um Sprache angemessen zu verwenden.

Es ist durchaus möglich, strategische Kompetenz so zu verstehen, dass sie sowohl Bewertung und Planung als auch Kompensation und Korrektur umfasst. Daraus ergibt sich, dass die strategische Kompetenz, je nach den Umständen, vom Lerner bewusst eingesetzt wird oder im Unterbewussten wirkt. Die Anstrengung, die damit verbunden ist, eine Fremdsprache zu lernen, und die im Vergleich zum Muttersprachler geringe zielsprachliche Kompetenz legen nahe, dass die strategische Kompetenz vom Fremdsprachenlerner bewusster eingesetzt wird als vom Muttersprachler.

Ausgehend von den Auswirkungen des bisher Gesagten für den Fremdsprachenunterricht und das Testen kann man zwei Typen von Aufgaben unterscheiden. Aufgaben zum Sprechen verlangen eine unmittelbare Reaktion und zwingen somit den Lerner, Formen routinemäßiger Kommunikation zu erlernen und sich Verfahren anzueignen, um mit Zusammenbrüchen in der Kommunikation umgehen zu können. Aufgaben zum Schreiben hingegen können bewusstes

Planen beinhalten, ja dies sogar erforderlich machen, da sie keine unmittelbare Reaktion erfordern. Die Fähigkeit zur Selbstbewertung sowie die vom Lerner erworbene Kontrolle über den eigenen Lernprozess hängen unmittelbar mit dem Erwerb der strategischen Kompetenz zusammen. Was Tests betrifft, könnte es möglich sein, in einem Test zum Sprechen auch Aufgaben anzubieten, die den Lerner vermutlich dazu bringen, Kompensationsstrategien einzusetzen. Der Prüfer würde dann bewerten, wie gut solche Strategien eingesetzt wurden. Dabei gibt es jedoch das Problem, dass die besten Kompensationsstrategien diejenigen sind, die nicht bemerkt werden! Es wäre vermutlich einfacher, das Element Planung der strategischen Kompetenz zu testen, indem man die Notwendigkeit zur Planung so in eine Aufgabe zum Schreiben aufnimmt, dass das Planen als gesondertes Element der Aufgabenerfüllung bewertet werden kann.

i) Änderungen am Bachman Modell

1996 stellten Bachman und Palmer eine modifizierte Fassung ihres Modells der kommunikativen Sprachkompetenz vor (siehe Abbildung 1). Das erste Merkmal des Modells, auf das wir eingehen möchten, ist das Sprachwissen. Dies beinhaltet alle Aspekte dessen, was man als Wissen über das formale System der Sprache bezeichnen könnte, gekoppelt mit den Besonderheiten des Sprachgebrauchs.

Sprachwissen

Strukturelles Wissen		Pragmatisches Wissen		
Grammatisches Wissen	Textuelles Wissen	Lexikalisches Wissen[1]	Funktionales Wissen	Soziolinguistisches Wissen
Syntax	Rhetorik	Semantik	kognitiv/affektiv	Konventionen des Sprachgebrauchs
Morphologie	Kohäsion	Denotation	manipulativ	
Orthographie/		Konnotation	heuristisch	Dialekt/Sprachvarietäten
Phonologie			imaginativ	Register
				Natürlichkeit

Abbildung 1 (Bachman und Palmer, 1996)

Sprachwissen ist eine offensichtliche und wesentliche Voraussetzung für den Gebrauch einer Sprache. Deren erfolgreiche Anwendung wird aber entscheidend durch die Interaktion mit anderen mentalen Prozessen beeinflusst, namentlich mit Wissensschemata und affektiven Schemata. Wissensschemata beziehen sich auf das Wissen von und die Erfahrungen mit der Welt, wohingegen sich affektive Schemata auf das emotionale Gedächtnis beziehen. Das Modell unterstellt, dass diese drei mentalen Bereiche durch einen metakognitiven Prozess aktiviert und nutzbar gemacht werden. Dieser verläuft in drei Phasen: Bewertungsstrategien, Planungsstrategien und Zielsetzungsstrategien.

In einem Gespräch mit einem Freund z. B. kann der Sprecher auf der Grundlage seiner kognitiven Schemata eine Bewertung des Themas vornehmen. Von dieser ausgehend, bewertet und plant er die Sprache, die ihm zum Erreichen des selbst gesetzten kommunikativen Ziels zur Verfügung steht. Wenn nach abgeschlossener Einleitung des Gesprächs neue Informationen gegeben werden, wie z. B. über die Beförderung des Freundes, wird dies eine weitere Bewertung, Planung und Zielsetzung sowie die Verwendung anderer sprachlicher Mittel notwendig machen. Das Modell erweist sich somit als dynamisch. Auf der einen Seite beinhaltet es einen kontinuierlichen Austausch zwischen den metakognitiven Prozessen und dem Sprachwissen sowie den affektiven und den Wissensschemata des Sprachverwenders, auf der anderen Seite umfasst es einen Austausch

[1] Anmerkung (TestDaF-Institut): Lexik wird im Originaltext von Bachman & Palmer dem Grammatischen Wissen zugeordnet.

zwischen dem Sprachverwender und dem Kontext der Sprachverwendung.

Für die Entwickler von Fremdsprachentests ist ein Modell der kommunikativen Kompetenz deshalb von Nutzen, weil es die Grundlage für die Definition der zu prüfenden Kompetenzbereiche bildet. Nur eine klare Vorstellung von dem, was zu testen ist, ermöglicht es festzustellen, ob ein Test als valide gelten kann oder nicht. Darüber hinaus können auf diese Weise so nützliche Werkzeuge wie Checklisten für die Inhalte eines Tests erstellt werden.

Das übergeordnete Ziel aller Fremdsprachentests besteht darin, Beispiele der sprachlichen Fähigkeiten und Fertigkeiten eines Kandidaten so zu erfassen, dass sie ein adäquates Abbild seiner Fremdsprachenbeherrschung außerhalb einer Testsituation darstellen. Ein Test, der diese Aufgabe angemessen erfüllt, ist ein valider Test. Was genau Validität ist und wie sie erreicht werden kann, war – und ist bis heute – bei Fremdsprachentests ein umstrittener Punkt.

5. Validität

a) Verfahren zur Ermittlung von Validität

Üblicherweise wird der Begriff Validität definiert als das Ausmaß, in dem ein Test das misst, was er messen soll (Pratt, 1980; Popham, 1981; Priestly, 1982; Carroll und Hall, 1985), oder das Ausmaß, in dem er Informationen bereitstellt, die relevant sind für die Entscheidung, die auf der Basis des Tests getroffen werden muss (Thorndike und Hagen, 1977). Die Validität eines Tests kann auf unterschiedliche Weise nachgewiesen werden: durch einen Vergleich des Testinhalts mit dem Curriculum, auf das sich der Test bezieht; durch einen Vergleich der Testergebnisse mit den Ergebnissen in einem anderen etablierten Test, der die gleiche Fertigkeit oder Eigenschaft misst; durch einen Vergleich der Leistungen in dem Test mit dem jeweiligen Erfolg in einem bestimmten Bereich; oder durch den Versuch, Beweise für die Annahme zu finden, dass die getesteten Fertigkeiten in der Tat das Konstrukt der Sprachkompetenz widerspiegeln, das dem Test zugrunde liegt.

b) Aspekte der Validität und wie man sie erreicht

Es sind unterschiedliche Verfahrensweisen vorgeschlagen worden, um Validität zu erreichen. Einige Jahre lang beschäftigte sich die Diskussion dieser Frage damit, die unterschiedlichen Aspekte der Validität zu definieren. Die „Standards for Educational and Psychological Tests“ (1974), mit denen Popham (1981), Cronbach (1970) sowie Thorndike und Hagen (1977) übereinstimmen, definierten drei Arten von Validität: inhaltliche Validität, kriterienbezogene Validität und Konstruktvalidität. Darüber hinaus wurde eine vierte Art von Validität vorgeschlagen: die Augenscheinvalidität.

i) Inhaltliche Validität

Die inhaltliche Validität, die gelegentlich auch als das Prinzip der „inclusiveness“ (McCormick und James, 1983) bezeichnet wird, betrifft das Ausmaß, in dem ein Test die Inhalte abdeckt, die er abdecken soll. Dieser Aspekt der Validität ist äußerst wichtig bei Kursabschlusstests oder Lernfortschrittstests im Unterricht und in der Schule (Deale, 1975). Bei Kursabschlusstests wird der Inhalt durch das Curriculum oder die Lehrwerke festgelegt. Die inhaltliche Validität kann z. B. durch einen Vergleich des Curriculums mit dem Test festgestellt werden. Theoretisch ist die inhaltliche Validität umso höher, je stärker beide übereinstimmen. Wenn die Sprachkompetenz unabhängig vom Unterricht festgestellt werden soll, ist es in der Regel der Testentwickler, der die Inhalte des Tests definiert. Moller (1982: 37) erläutert dies näher:

> „Content validity, together with reliability, will ensure that a test adequately reflects the objectives and linguistic content laid down in the syllabus. In the case of a proficiency test, however, the test constructors themselves decide the "syllabus" and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe. Thus the evaluator looking for content validity is really assessing the test constructor's definition of proficiency."

Allerdings trifft das gleiche Argument auch auf Kursabschlusstests zu. Jemand entwirft irgendwo ein Curriculum, auf dessen Basis Lehrbuchautoren ihre Lehrwerke verfassen. Die Entscheidung über die inhaltliche Angemessenheit des Curriculums und das Ausmaß, in dem die Lehrwerke dieses widerspiegeln, sind meistens subjektiv. So kann sich ein Lehrwerk auf den Threshold Level des Europarats berufen. Als Lehrwerk kann es mehr oder weniger angemessen den Treshold Level umsetzen. Institutionen verwenden dann das Lehrwerk und erstellen möglicherweise sogar begleitende Tests. Diese Tests mögen valide sein im Hinblick auf das Lehrwerk, aber es gibt keine Garantie, dass sie wirklich valide sind im Hinblick auf den Threshold Level, und übrigens auch keine Garantie, dass der Threshold Level selbst valide ist.

ii) Kriterienbezogene Validität

Die kriterienbezogene Validität betrifft das Ausmaß, in dem ein Test voraussagekräftig und/oder deckungsgleich ist, d. h., inwieweit er das Gleiche misst wie ein Test, der sich schon als angemessen erwiesen hat. Das gängigste Verfahren, um die kriterienbezogene Validität zu überprüfen, ist die Korrelation.

Um die Übereinstimmungsvalidität eines Tests nachzuweisen, werden üblicherweise die Ergebnisse der Kandidaten in dem Test, der untersucht werden soll, mit ihren Ergebnissen in einem anderen Test korreliert, von dem angenommen wird, dass er das Gleiche testet. Das Problem besteht darin, schlüssig zu beweisen, dass die beiden Tests tatsächlich das Gleiche auf die gleiche Art und Weise messen. Dies wurde lange als Schwierigkeit angesehen und neuere Forschungen im Bereich der Merkmals- und Methodeneffekte (Bachman und Palmer, 1983; Shoamy, 1984) haben sehr deutlich gezeigt, dass es keine Garantie dafür gibt, dass Methoden äquivalent sind. Wenn neue Aufgabentypen entwickelt werden, ist es deshalb sehr fraglich, ob die üblichen Verfahren zur Bestimmung der Übereinstimmungsvalidität angewendet werden können.

In manchen Fällen kann die Übereinstimmungsvalidität durch Korrelation der Testergebnisse mit den Noten der Lehrer hergestellt werden (Chaplen, 1970). Ingram (1974) geht sogar so weit zu behaupten, dass ein Vergleich mit den Noten der Lehrer das beste Verfahren sei, um die Übereinstimmungsvalidität eines Tests festzustellen. Es ist jedoch problematisch, Kriterien zu entwickeln, an die sich die Lehrer halten müssen (Moller, 1982: 52), und die Interpretation dieser Kriterien durch die Lehrer zu standardisieren.

Wenn schon das Herstellen von Übereinstimmungsvalidität mit Schwierigkeiten behaftet ist, so gilt dies auch für die Vorhersagevalidität. Viele Studien bezogen sich auf den akademischen Kontext, da Lerner dort am häufigsten Sprachtests ablegen müssen und eine ausreichende Menge an Daten für eine Untersuchung vorliegt. Moller (1982) begutachtet eine Reihe von Untersuchungen zur Vorhersagevalidität und stellt fest, dass die meisten Untersuchungen nicht-linguistische Kriterien verwenden, wie z. B. Durchschnittsnoten. Die Ergebnisse von Untersuchungen zur Vorhersagevalidität können durch viele Faktoren beeinflusst werden, wie die Länge des Kontakts der Lerner mit der Fremdsprache, ihre Fähigkeit und Bereitschaft, Fortschritte zu machen, das Ausmaß, in dem die Lehrer bestimmte Aspekte der Sprache erklären usw. Darüber hinaus sind in vielen Untersuchungen zur Vorhersagevalidität die verwendeten Daten schon dadurch verfälscht, dass nur Daten von Studierenden erhoben werden können, deren Ergebnisse in den Englischtests es ihnen erlauben, ein Studium in einem englischsprachigen Land aufzunehmen. Alle diejenigen, die keine ausreichenden Ergebnisse erzielt haben, sind von vornherein ausgeschlossen. Ob einige von ihnen im Studium erfolgreich gewesen wären, bleibt offen.

iii) Konstruktvalidität

Um den Nachweis der Konstruktvalidität zu erbringen, muss bewiesen werden, dass der Test die psychologischen Konstrukte misst, die er zu messen vorgibt. Dem liegt die Annahme zugrunde, dass Tests dazu gedacht sind, uns Informationen über ein Phänomen in der realen Welt, ein Merkmal oder ein Verhalten zu liefern. Tests sind im Allgemeinen ein indirektes Instrument zur Feststellung, in welchem Maß ein Individuum ein theoretisch angenommenes Konstrukt oder Merkmal besitzt.

Das Verfahren zur Konstruktvalidierung kann nach Walsh und Betz (1985) in drei Phasen unterteilt werden:

> „First, the construct of interest is carefully defined and the hypotheses regarding the nature and extent of its relationships to other variables are postulated. Second, an instrument designed to measure that construct is developed. Third, after the degree to which the test is reliable has been examined, studies examining the relationship of the test to other variables (as formulated in the hypotheses about the construct of interest) are undertaken."

Um die Konstruktvalidität zu untersuchen, gibt es unterschiedliche statistische Methoden. Die in letzter Zeit häufig eingesetzte Faktorenanalyse der Interkorrelation der Subtests oder Items stellt fest, wie viele Dimensionen oder Merkmale notwendig sind, um Testergebnisse zusammenzufassen oder zu erklären. Wenn z. B. ein Test dazu dienen soll, mehr als ein Merkmal zu messen, die Faktorenanalyse aber nur einen „allgemeinen" Faktor feststellt, dann kann der Test, zumindest aus dieser Sicht, keine Konstruktvalidität besitzen.

Im Bereich der Sprachtests gab es bis in die späten 70er Jahre verhältnismäßig wenig Interesse an Untersuchungen zur Konstruktvalidität, erst danach tauchten in der Forschung, besonders in den USA, eine ganze Reihe von Untersuchungen auf. Die meisten dieser Untersuchungen können nach Weir (1984: 65) folgendermaßen gesehen werden:

> „… principally as the a posteriori statistical validation of whether a test has measured a construct which has a reality independent of other constructs. The concern is much more with the a posteriori relationship between a test and the psychological abilities, traits, constructs, it has measured than with what it is that it should have elicited in the first place."

So wichtig die Post-Validierung eines Tests auch sein mag, so wichtig ist es auch, von vornherein die Angemessenheit der Testinhalte sicher zu stellen. Das heißt, bevor nachträglich festgestellt werden kann, dass etwas getestet wurde, muss es zuvor eine klare Definition dessen geben, was getestet werden soll. Ebel (1983) weist darauf hin, dass der Hauptgrund für die Schwierigkeiten bei der Testvalidierung darin liegt, dass die Notwendigkeit empirischer Daten für die Validierung überbetont wird und die enorme Wichtigkeit von expliziten verbalen Definitionen dessen, was der Test messen soll, nicht gesehen wird.

Eine stärkere Fokussierung auf eine Validierung von Beginn an (a priori) führt unausweichlich zu einer Überschneidung zwischen Inhalts- und Konstruktvalidität.

iv) Augenscheinvalidität

Obwohl die Augenscheinvalidität wichtig ist, zählt sie wegen der fraglichen Rolle, die sie bei der Validierung spielt, normalerweise nicht zu den drei Haupttypen der Validität.

Die Augenscheinvalidität, d. h. das Ausmaß, in dem ein Test in den Augen der Kandidaten, der Lehrer und der Auftraggeber das Richtige testet, spielte in der Diskussion zur Validität von Sprachtests zwar auch eine Rolle, es gibt aber kein allgemein anerkanntes Verfahren, um festzustellen, ob ein Test Augenscheinvalidität besitzt. Dies bewog einige Experten dazu, der Augenscheinvalidität keinen Platz in der Diskussion über Testvalidität einzuräumen (Bachman et al., 1981).

Obwohl Stevenson (1985) im Bereich des Testens den Trend zu mehr performanzorientierten Formaten unterstützt hat, warnt er, dass die Augenscheinvalidität für messtechnisch naive Beobachter nur den Anschein einer Validität biete. Sie führe zu einer psychometrisch undifferenzierten Selbstsicherheit, die es erlaubt, einen Test einfach nur anzusehen und ohne weitere Prüfung zu behaupten, dass „ich einen validen Test erkenne, wenn ich einen sehe."

Auch Stanley und Hopkins (1972: 105) weisen darauf hin, dass die Augenscheinvalidität auf einer sehr naiven und oberflächlichen Beurteilung beruht und dass es gefährlich ist, ihr zu viel Bedeutung beizumessen. Während ein Test mit einer guten inhaltlichen Validität normalerweise auch Augenscheinvalidität aufweist, ist dies umgekehrt nicht notwendigerweise der Fall.

Bachman (1990: 285-289) gibt einen Überblick über die Fachdiskussionen seit 1947 (Mosier). Er stellt fest, dass die meisten Ausführungen zum Begriff Augenscheinvalidität (oder, wie er es nennt, „test appeal") negativ waren und schlägt ein Moratorium für diesen Begriff vor. Für ihn fand die „endgültige Beisetzung" dadurch statt, dass der Begriff in der letzten Ausgabe (1985) der „Standards for Educational and Psychological Testing" (APA – American Psychological Association) mit keinem Wort erwähnt wurde. Bachman stellt fest:

> „… the „bottom line" in any language testing situation, in a very practical sense, is whether test takers will take the test seriously enough to try their best, and whether test users will accept the test and find it useful. For these reasons, test appearance is a very important consideration in test use." (1990: 288)

Abschließend kann daher festgehalten werden, dass ein Test ohne Zweifel so aussehen sollte, als prüfe er die richtigen Dinge auf die richtige Weise. Als ebenso unstrittig muss jedoch gelten, dass ein Validitätsnachweis nicht primär über die Augenscheinvalidität angegangen werden kann, da sich letztere automatisch ergeben sollte, wenn man sorgfältig darauf achtet, dass der Test inhaltlich valide ist.

c) Neuere Ansichten zur Testvalidität

In letzter Zeit änderte sich die Auffassung, dass man Validität in unterschiedlich bezeichnete einzelne Arten unterteilen könne, und man sieht nun Validität als einen andauernden Prozess und integralen Bestandteil der Validierung. Diese Entwicklung begann in den frühen 80er Jahren, als Cronbach zum Beispiel die Meinung äußerte, dass „jede Form der Validierung als ein Ganzes zu sehen ist" (1980: 99). Seine Auffassung spiegelt wider, dass nun die Betonung auf die Konstruktvalidität als wichtigsten Bestandteil eines umfassenden Begriffs von Validität gelegt wurde.

Heute wird deshalb Validität als ein umfassendes Konzept gesehen. Statt von verschiedenen Arten der Validität zu sprechen, wird in der Fachliteratur jetzt von Kategorien des Validitätsnachweises (siehe APA Standards 1985) gesprochen. Bachman (1990) teilt diese Auffassung und definiert Validierung als „Sammlung von Validitätsnachweisen". Der Prozess der Validierung kann daher als eine Form der wissenschaftlichen Untersuchung angesehen werden, die die systematische Sammlung von Informationen (d. h. Validitätsnachweisen) erfordert, und die im Gegenzug eine angemessene Interpretation der Testergebnisse durch Prüfungsabnehmer ermöglicht (d. h. die Auswirkungen der Testanwendung).

Mit seinen Veröffentlichungen hatte Messick (1989) erheblich dazu beigetragen, diese Auffassung von Validität durchzusetzen. Er betont, dass die Konstruktvalidität alleine nicht ausreicht, obwohl sie ein äußerst wichtiger Nachweis für die Interpretation der Testergebisse ist. Er schlägt eine progressive Matrix vor, in der die Konstruktvalidität die Hauptkomponente in jeder Zelle bildet, die aber auch andere Komponenten enthält, wie die Berechtigung des Testens und die Auswirkungen, die durch den Einsatz eines bestimmten Tests entstehen.

i) Die Matrix von Messick (1989: 20)

	Funktion des Ergebnisses	
Grundlage der Berechtigung	*Interpretation des Tests*	*Verwendung des Tests*
Auf der Basis von Nachweisen	Konstruktvalidität	Konstruktvalidität + Relevanz/Nützlichkeit
Auf der Basis von Auswirkungen	Folgerungen hinsichtlich Güte	Soziale Auswirkungen

Nach Messick sollte die Konstruktvalidität nur als ein Nachweis (in Verbindung mit Relevanz, Nützlichkeit, Folgerungen hinsichtlich Güte und sozialen Auswirkungen) bei der Verwendung des Tests sowie der Interpretation der Tests angesehen werden. Die inhaltsbezogene Validität z. B. wird jetzt als eine Reihe von Nachweisen gesehen, die die Konstruktvalidität unterstützen und die die notwendigen Informationen für eine angemessene Interpretation der Testergebnisse liefern.

Die Matrix von Messick wurde bei Tests, die für den Echteinsatz vorgesehen sind, noch nicht angewendet (nur partielle Anwendungen wurden bisher publiziert), was in nicht geringem Maße daran liegt, dass es schwierig ist zu verstehen, wie Messick sich diese Anwendungen genau vorgestellt hat. Wie im folgenden Abschnitt dargelegt, hat sich die Auffassung von Messick durchgesetzt und die heutige Testvalidierung maßgeblich mitbestimmt, d. h. die Auffassung von Validität als einem umfassenden Konzept, das die Testentwickler zwingt, eine Reihe von Nachweisen zu erbringen. Außerdem hatte der Begriff „Auswirkungsvalidität des Tests“ weit reichende Folgen, denn er zieht die Testentwickler zur Verantwortung für die beabsichtigten – und unbeabsichtigten – Auswirkungen eines Tests und zwingt sie, diesen Beachtung zu schenken.

d) Praxisorientierte Rahmenmodelle für die Testvalidierung

Der neueste Trend in der Validitätstheorie versucht, über die rein theoretischen und oft nicht praxisorientierten Validitätsmodelle hinauszugehen und eine praxisnähere und anwendbare Beschreibung der Validität zu liefern. Diese Beschreibungen erfolgen üblicherweise in Form von Rahmenvorgaben, da diese nach Meinung ihrer Anhänger eine theoretisch fundierte Begründung sowohl für die Testentwicklung als auch für die Validierung liefern. Die einflussreichsten dieser Rahmenvorgaben wurden von Mislevy (in den USA) und Weir (in Europa) vorgestellt.

i) Nachweiszentrierte Testentwicklung: Mislevy

Robert Mislevy und seine Mitarbeiter präsentierten in den späten 90ern und in den ersten Jahren des neuen Jahrtausends in einer Reihe von Veröffentlichungen ein Rahmenmodell für die Testentwicklung, das durch eine Unterteilung in vier Phasen gekennzeichnet ist:

Phase	Titel	Kurze Beschreibung
1	Analyse des Wissensgebiets	Ein weiter Bereich, der das Sammeln von Informationen zum Wissensgebiet aus verschiedenen Quellen beinhaltet. Dies bezieht sich weitgehend auf den Bereich der Bedarfsanalyse.
2	Modellierung des Wissensgebiets	Dies bezieht sich im weitesten Sinne auf die Beschreibung dessen, was Mislevy (2003: 6) die „evidentiary relationships" (nachweisbasierte Zusammenhänge) genannt hat, also auf die Zusammenhänge zwischen den wichtigsten Paradigmen, die das Wissensgebiet definieren (Aussagen, Nachweis und Aufgabenparadigmen).
3	Konzeptioneller Rahmen für den Test	Dies bezieht sich aus unterschiedlichen Perspektiven auf die Zielbeschreibungen des einsatzbereiten Tests oder Beurteilungsverfahrens. *Modelle von Lernern:* Merkmale der Kandidaten, die für die Leistungsmessung relevant sind *Modelle von Aufgaben:* Verfahren, um Sprachproduktion auszulösen *Modelle von Nachweisen:* Bewertung von Aufgaben und Tests *Modelle von Testproduktion:* Wie ein Test aus ausgewählten Aufgaben zusammengestellt wird *Modelle von Präsentationen:* Wie die Aufgaben dargeboten, die Interaktionen geleitet und die Leistung erfasst werden sollen
4	Beurteilung des Testeinsatzes	Die Phase im Zyklus der Testentwicklung, in der ein Test zum Einsatz kommt.

Wie man der Tabelle oben entnehmen kann, ist die Analyse des Wissensgebiets (*engl. domain analysis*) die erste Phase der Entwicklung, in der Daten gesammelt werden. Hier versucht der Testentwickler, das Wissen und die Aufgaben zu beschreiben, die das jeweilige Gebiet unter verschiedenen Gesichtspunkten definieren. Darauf folgt die Modellierung des Wissensgebiets (*engl. domain modelling*), die McNamara (2003: 9) für die wesentlichste Phase im Entwicklungsprozess hält. Sie ist in drei Stufen untergliedert: Aussagen, Nachweis und Aufgaben. Damit ist gemeint, dass das Konstrukt dadurch definiert wird, dass man zuerst die Aussagen klärt, die man hofft, über die Kandidaten machen zu können, und dann die Art des Nachweises nennt, der notwendig ist, um diese Aussagen zu unterstützen. Zum Schluss (d. h. bei den Aufgabenparadigmen) muss der Testentwickler entscheiden, wie dieser Nachweis erbracht werden soll, d. h., er muss sich für den effizientesten Weg entscheiden, um die Fähigkeiten/Fertigkeiten beobachten zu können, die gemessen werden sollen.

Modellierung des Wissengebiets (Begründung)		
AUSSAGEN	NACHWEIS	AUFGABEN
Die Merkmale der Lerner und in welchen Aspekten der Sprachfähigkeit sie sich äußern.	Die Merkmale dessen, was Lerner sagen und tun – was müsste ein Lerner sagen oder tun, um eine Grundlage für Aussagen über ihn zu schaffen? (Wie identifiziert man Nachweise in den Arbeiten der Lerner – anhand welcher Kriterien?)	Die Arten von Situationen, die es möglich machen, diesen Nachweis zu erhalten und die die konstruktirrelevante Varianz minimieren.

(McNamara, 2003: 9)

In der Phase 3 (konzeptioneller Rahmen für den Test) erstellt der Testentwickler detaillierte Testbeschreibungen (d. h. die Beschreibung der Prüfungsziele, des Testformats, der Item- und Aufgabentypen und der Bewertungsverfahren), die die Grundlage für den endgültigen Test bilden. Diese Testbeschreibungen sollten unbedingt in den frühen Phasen der Testentwicklung erstellt werden, um sicherzustellen, dass der endgültige Test bei seinem Einsatz die Definition des Konstrukts widerspiegelt (Aussagen, Nachweis und Aufgaben, wie in Phase 2 des Rahmens ausgeführt). Die letzte Phase der Rahmenvorgaben (Beurteilung des Testeinsatzes) ist selbsterklärend. Das Ergebnis all dieser Überlegungen, Planungen und Beschreibungen ist ein einsatzbereites Leistungsmessungsverfahren.

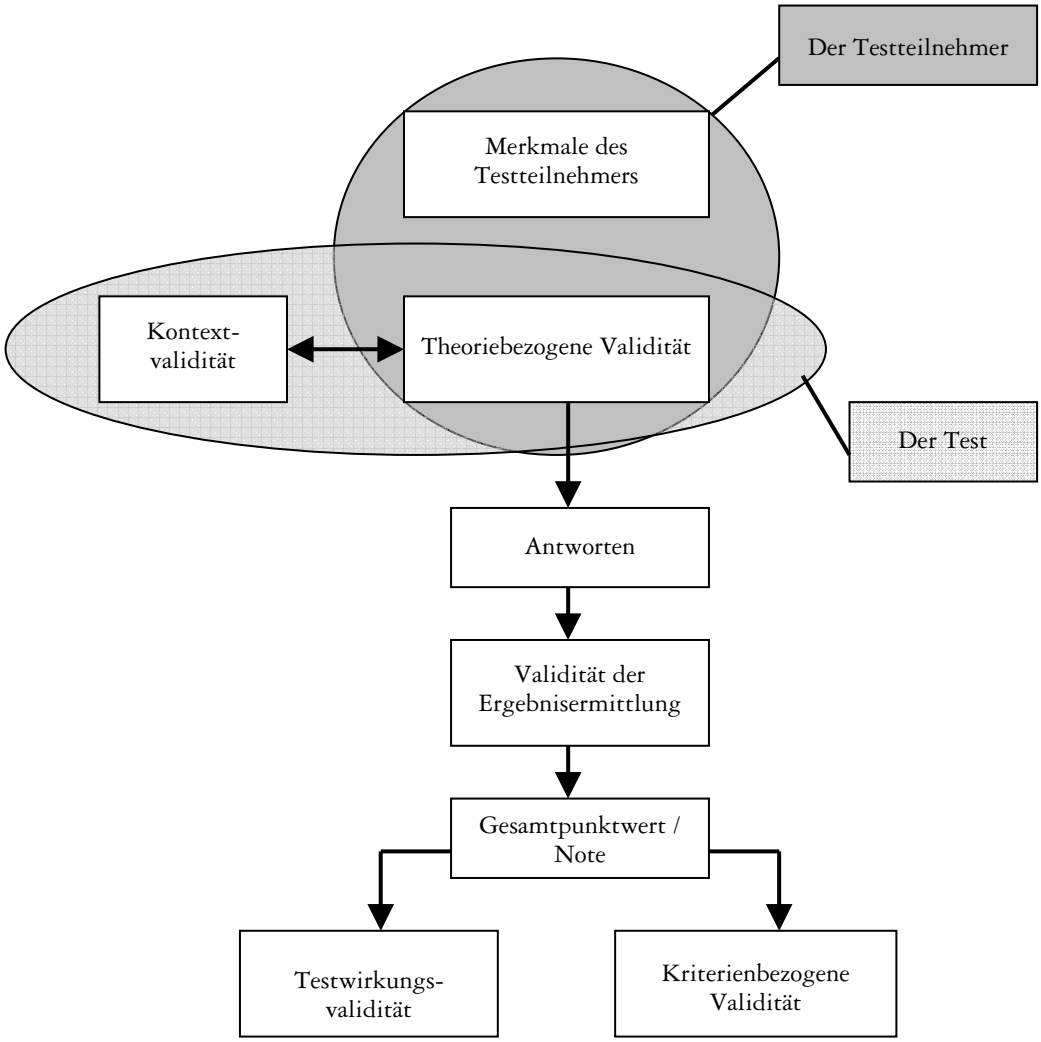
Dieser Rahmen wurde als Teil der theoretischen Begründung für die letzte Version des TOEFL Tests (Test of English as a Foreign Language) von ETS (Educational Testing Service) in den USA eingesetzt. Es ist jedoch nicht klar, inwieweit Mislevys Vorstellungen im Bereich der Testentwicklung tatsächlich in die Praxis umgesetzt werden. Seine Auffassungen sind komplex und scheinen besser geeignet für Tests mit sehr hohen Teilnehmerzahlen, die bei der Testdurchführung, den statistischen Analysen und den Rückmeldeverfahren sehr auf Technik setzen. Ungeachtet dessen bietet der Rahmen eine interessante Sicht auf den Zusammenhang zwischen theoretischer Begründung eines Tests und der Umsetzung in den Testbeschreibungen. Dadurch könnte er sowohl für Testentwickler als auch für kritische Begutachter von Interesse sein.

Zur gleichen Zeit als Mislevy und seine Mitarbeiter ihren Rahmen entwickelten, erarbeiteten Cyril Weir und sein Kollege Barry O'Sullivan am CRTEC (Centre for Research in Testing, Evaluation and Curriculum in ELT) in London ihren Rahmen zur Validierung von Tests, wiederum mit einem Ansatz, der sich auf Nachweise stützt, jedoch mit einem expliziteren Verweis auf die sozio-kognitiven Aspekte des Sprachgebrauchs.

ii) Rahmenmodell zur Validierung von Tests (Weir)

Weir (2005) schlägt ein umfassendes Rahmenmodell vor, das seiner Ansicht nach die Grundlage für jede Art von Testentwicklung und Validierung bilden kann. Im folgenden Überblick sind einige Elemente daraus aufgelistet, die Testentwickler berücksichtigen sollten. Weir (2005:48) betont, dass Testentwickler versuchen müssen, auf jede der folgenden Fragen einzugehen:

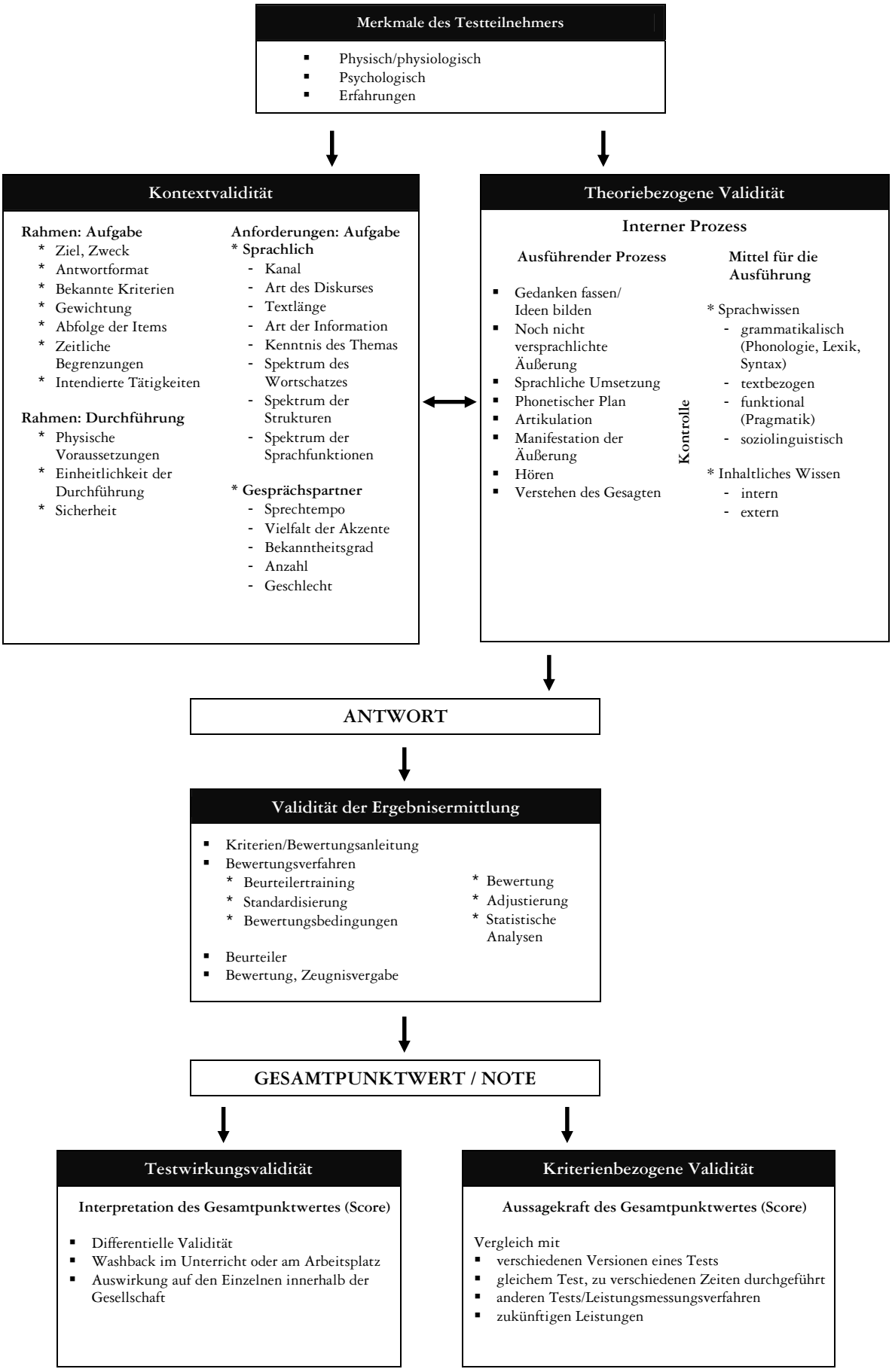
- Wie werden die physischen, physiologischen und psychologischen Merkmale der Lerner und die Merkmale ihrer Erfahrungen im Test berücksichtigt? (Testteilnehmer)
- Sind die Merkmale der Testaufgabe(n) und der Testdurchführung fair gegenüber den Kandidaten, die den Test ablegen? (Kontextvalidität)
- Sind die kognitiven Prozesse, die zur Lösung der Aufgaben erforderlich sind, angemessen? (Theoriebezogene Validität)
- Inwieweit kann man sich auf die im Test erzielten Ergebnisse verlassen? (Validität der Ergebnisermittlung)
- Welche Wirkung hat der Test auf alle an ihm Beteiligten und von ihm Betroffenen? (Testwirkungsvalidität)
- Welchen externen Nachweis gibt es außerhalb der Testergebnisse selbst, dass der Test gute Arbeit leistet? (Kriterienbezogene Validität)



Zum Zweck der Beschreibung sind die einzelnen Elemente in diesem Modell so dargestellt, als seien sie unabhängig voneinander. Weir weist aber darauf hin, dass es eine symbioseartige Verbindung zwischen Kontextvalidität, theoriebezogener Validität und Validität der Ergebnisermittlung gibt, die er alle als Teil der Konstruktvalidität sieht. Beispiele für diese symbioseartige Verbindung:

- Entscheidungen, die im Hinblick auf Parameter des Aufgabenkontexts gefällt werden, beeinflussen die Vorgänge, die bei der Lösung der Aufgaben stattfinden.
- Gibt man den Kandidaten die Bewertungskriterien vorher bekannt, hat dies einen Einfluss auf die Prozesse, die bei der Planung und Ausführung der Aufgabe stattfinden.

Weir schlägt vier solcher Rahmen vor, einen für jede Fertigkeit. Um zu zeigen, wie solche Rahmenvorgaben aussehen können und wie man sie anwenden kann, wird im Folgenden ein Beispiel für die Fertigkeit Sprechen gegeben.


Merkmale des Testteilnehmers
Physisch/physiologisch
Psychologisch
Erfahrungen
Kontextvalidität
Rahmen: Aufgabe
* Ziel, Zweck
* Antwortformat
* Bekannte Kriterien
* Gewichtung
* Abfolge der Items
* Zeitliche Begrenzungen
* Intendierte Tätigkeiten
Rahmen: Durchführung
* Physische Voraussetzungen
* Einheitlichkeit der Durchführung
* Sicherheit
Anforderungen: Aufgabe
* Sprachlich
- Kanal
- Art des Diskurses
- Textlänge
- Art der Information
- Kenntnis des Themas
- Spektrum des Wortschatzes
- Spektrum der Strukturen
- Spektrum der Sprachfunktionen
* Gesprächspartner
- Sprechtempo
- Vielfalt der Akzente
- Bekanntheitsgrad
- Anzahl
- Geschlecht
Theoriebezogene Validität
Interner Prozess
Ausführender Prozess
Gedanken fassen/ Ideen bilden
Noch nicht versprachlichte Äußerung
Sprachliche Umsetzung
Phonetischer Plan
Artikulation
Manifestation der Äußerung
Hören
Verstehen des Gesagten
Kontrolle
Mittel für die Ausführung
* Sprachwissen
- grammatikalisch (Phonologie, Lexik, Syntax)
- textbezogen
- funktional (Pragmatik)
- soziolinguistisch
* Inhaltliches Wissen
- intern
- extern
ANTWORT
Validität der Ergebnisermittlung
Kriterien/Bewertungsanleitung
Bewertungsverfahren
* Beurteilertraining
* Standardisierung
* Bewertungsbedingungen
* Bewertung
* Adjustierung
* Statistische Analysen
Beurteiler
Bewertung, Zeugnisvergabe
GESAMTPUNKTWERT / NOTE
Testwirkungsvalidität
Interpretation des Gesamtpunktwertes (Score)
Differentielle Validität
Washback im Unterricht oder am Arbeitsplatz
Auswirkung auf den Einzelnen innerhalb der Gesellschaft
Kriterienbezogene Validität
Aussagekraft des Gesamtpunktwertes (Score)
Vergleich mit
verschiedenen Versionen eines Tests
gleichem Test, zu verschiedenen Zeiten durchgeführt
anderen Tests/Leistungsmessungsverfahren
zukünftigen Leistungen

Weir (2005:46)

Die verschiedenen Parameter werden in den folgenden Tabellen kurz beschrieben.

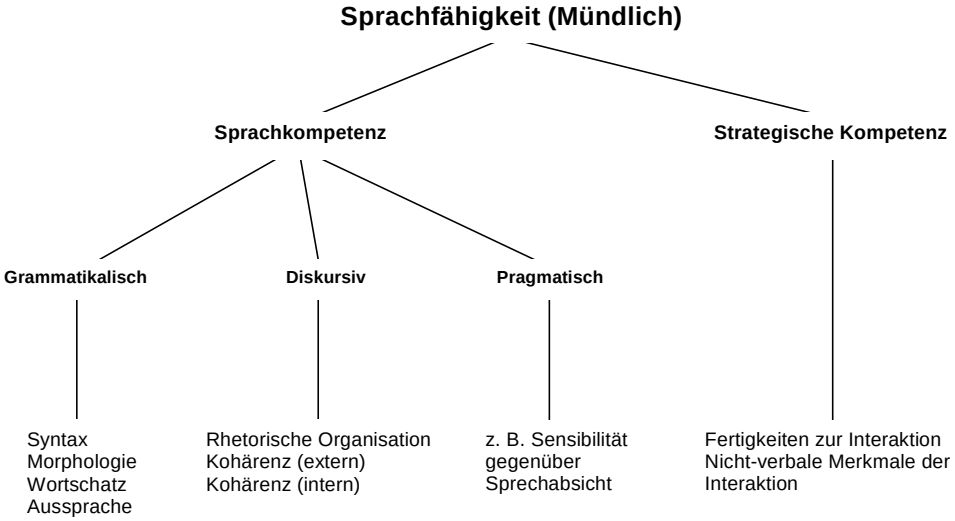
MERKMALE DES TESTTEILNEHMERS	
Physisch/Physiologisch	
Kurzzeitiges Unwohlsein	*Zahnschmerzen, Erkältung usw.*
Längerfristige Behinderungen	*Sprach- Hör-, Sehstörungen (z .B. Lesestörung)*
Alter	Angemessenheit der Texte, der Themen usw. Anforderungen der Aufgabe (Zeit, kognitive Anforderung usw.)
Geschlecht	Angemessenheit der Texte, der Themen usw.
Psychologisch	
Gedächtnis	Bezieht sich auf die Art der Aufgabe, auch auf physische Merkmale.
Persönlichkeit	Bezieht sich beim Sprechen primär auf die Art der Aufgabe (z. B. Anzahl der Teilnehmer – Einzel-, Paar- oder Gruppenprüfung).
Kognitiver Stil	Der bevorzugte Lernstil z. B. kann die Aufgabenerfüllung beeinflussen.
Affektive Schemata	Wie der Kandidat auf die Aufgabe reagiert. Kann vom Testentwickler berücksichtigt werden, indem er den Zweck bzw. die Ziele der Aufgabe sorgfältig kontrolliert.
Konzentration	Hängt mit dem Alter zusammen und auch (besonders beim Hören und Lesen) mit der Länge und der Menge des Inputs.
Motivation	Hängt u. a. mit dem Thema der Aufgabe und dem Ziel der Aufgabe/des Tests zusammen.
Emotionaler Zustand	Ein Beispiel für eine unvorhersehbare Variable. Ist nicht einfach zu berücksichtigen, könnte aber so behandelt werden wie Motivation oder affektive Schemata.
Erfahrungen	
Bildung/Ausbildung	Kann formal oder informell sein und in einem Kontext stattgefunden haben, in dem die Zielsprache die Muttersprache oder die Zweitsprache war.
Prüfungsvorbereitung	Bezieht sich entweder auf einen Kurs, der auf eine spezielle Prüfung oder eine ähnliche Prüfung vorbereitet, oder ganz allgemein auf die Vorbereitung auf Prüfungen.
Prüfungserfahrung	Kann sich wiederum auf eine spezielle Prüfung, ähnliche Prüfungen oder auf Prüfungen allgemein beziehen.
Kommunikative Erfahrungen	Kann sich auf alles oben Genannte beziehen, z. B. darauf, dass sich die kommunikativen Erfahrungen auf den Unterricht beschränken, oder darauf, dass der Kandidat einige Zeit in einer zielsprachigen Gemeinschaft gelebt hat und an „echter“ Kommunikation in dieser Sprache teilgenommen hat.
Testsprache und Aufenthaltsort	Kann sich auf die schulische oder akademische Erziehung beziehen (d. h., wo sie stattgefunden hat) oder auf die kommunikativen Erfahrungen (z. B., ob die Sprache als Fremd- oder Zweitsprache gelernt wurde).

THEORIEBEZOGENE VALIDITÄT	
INTERNE PROZESSE (nach Levelt, 1989)	
Ideen generieren und ordnen	Eine Absicht denken, die auszudrückenden relevanten Informationen auswählen, um dieses Ziel zu erreichen, die Informationen für die Äußerungen ordnen, im Auge behalten, was zuvor gesagt wurde; aufmerksam das Gehörte und die eigenen Äußerungen verfolgen, indem man seine Wissensschemata aktiviert. Der Sprecher wird die Äußerungen kontrollieren, bevor er sie sprachlich umsetzt.
Noch nicht versprachlichte Äußerung	Ergebnis der Gedankenumsetzung, d. h. der oben genannten Vorgänge.
Sprachliche Umsetzung	Beinhaltet grammatische und phonologische Umsetzung im Hinblick auf lexikalische Form.
Phonetischer Plan	Eine innere Abbildung dessen, wie die geplante Äußerung ausgesprochen werden sollte; innere Rede.
Artikulation	Die Ausführung des phonetischen Plans durch die Muskeln des respiratorischen Systems, des Kehlkopfs sowie aller oberhalb liegenden Organe.
Manifestation der Äußerung	
Hören	Verstehen, was andere sagen oder was man selbst sagt, d. h. Sprechgeräusche als Worte oder Sätze mit Sinn zu interpretieren.
Verstehen des Gesagten	Zugang zu verschiedenen zur Verfügung stehenden Ressourcen, z. B. Lexikon, syntaktische Markierungen, Hintergrundwissen. Es wird eine Abbildung der Äußerung im Hinblick auf ihren phonologischen, morphologischen, syntaktischen und semantischen Aufbau gebildet. Betrifft sowohl die innere Rede als auch deren äußerliche Manifestation.
KONTROLLE	Eine Kontrolle der inneren Rede sowie deren Manifestation kann fortwährend stattfinden, obwohl dieser Kontrollfilter manchmal ausgeschaltet ist. Das System, das innere Ressourcen anzapft, um den Anforderungen der Ausführungsprozesse gerecht zu werden.

MITTEL FÜR DIE AUSFÜHRUNG	
Inhaltliches Wissen	
Intern	Das Vorwissen des Testteilnehmers hinsichtlich Thema oder kulturellem Inhalt (Hintergrundwissen).
Extern	Wissen, das durch den Test bzw. die Aufgabe(n) vermittelt wird.
Sprachwissen (alle Hinweise beziehen sich auf Buck, 2001)	
Grammatikalisch	Bedeutungsebene: beinhaltet Phonologie, Betonung, Intonation, Lexik und Syntax der gesprochenen Sprache.
Textbezogen	Längere Äußerungen oder interaktives Gespräch mit zwei oder mehr Sprechern: beinhaltet Kenntnis der Diskursmerkmale (Textzusammenhalt, rhetorische Schemata, Grammatik des Erzählens) und Wissen über die Struktur des ungeplanten Diskurses.
Funktional	Funktion oder kommunikative Wirkung einer Äußerung oder eines längeren Textes und das Interpretieren der beabsichtigten Bedeutung: beinhaltet zu verstehen, ob die Äußerung Ideen übermitteln, manipulieren oder der Kreativität Ausdruck geben soll; indirekte Sprechakte sowie deren pragmatische Auswirkungen verstehen.
Soziolinguistisch	Die Sprache in bestimmten soziokulturellen Rahmen und das Interpretieren der Äußerungen im Hinblick auf den Kontext der Situation: beinhaltet Kenntnis der angemessenen sprachlichen Formen sowie der Konventionen, die bestimmte Gruppen soziolinguistisch charakterisieren, und die Auswirkungen des Gebrauchs bzw. Nicht-Gebrauchs sprachlicher Formen wie Slang, idiomatischer Ausdrücke, Dialekte, kultureller Bezüge, Redewendungen, Ebenen der Formalität und Register.

Was die Ausführungsprozesse betrifft, sollte man sich unbedingt darüber im Klaren sein, dass der Testentwickler realistischerweise nur die erste Phase des Prozesses beeinflussen kann, z. B., indem er den Kandidaten die Kriterien verdeutlicht, die zur Leistungsbewertung herangezogen werden, oder indem er sicherstellt, dass die Anweisungen zur Lösung der Aufgaben eindeutig sind. Nur die Forschung kann zeigen, ob die kognitiven und metakognitiven Strategien, die der Kandidat einsetzt, um das Zusammenspiel der Prozesse und der Mittel zu kontrollieren, die ihm zur Ausführung der sprachlichen Handlung zur Verfügung stehen, die Prozesse widerspiegeln, die bei der Sprachverwendung in Realsituationen tatsächlich stattfinden und die der Testentwickler mit seinem Test abzubilden versucht.

Die Mittel, die dem Kandidaten zur Ausführung zur Verfügung stehen, schließen das in den Test eingebrachte inhaltliche Wissen ein (entweder als Hintergrundwissen oder als Wissen, das durch Informationen im Test erworben wird) sowie das Sprachwissen (das am häufigsten zitierte Modell ist das von Bachman, 1990 und Bachman/Palmer, 1996). Dieses Sprachwissen wird normalerweise in einem theoretischen Kompetenzmodell beschrieben, das die Grundlage für den Test bildet, wie z. B. das Modell von Saville und Hargreaves (1999: 45), auf das sich die Prüfungen von Cambridge ESOL beziehen:



KONTEXTVALIDITÄT	
Rahmen: Aufgabe	
Ziel, Zweck	Die Anforderungen der Aufgabe. Sie erlauben den Kandidaten, angemessene Strategien zu wählen und zu entscheiden, auf welche Informationen sie sich beim Textverstehen konzentrieren sollen und welche Strategien sie bei den produktiven Aufgaben aktivieren sollen. Unterstützt ***Zielsetzung*** und ***Kontrolle***.
Lösungsformat	Wie die Kandidaten die Aufgabe lösen sollen (z. B. Multiple-Choice im Gegensatz zu Kurzantwortaufgaben). Unterschiedliche Formate können sich unterschiedlich auf die Leistung auswirken.
Bekannte Kriterien	Den Kandidaten werden die Kriterien genannt, nach denen ihre Leistung bewertet wird. Das bedeutet, dass sie vor dem Test über die Bewertungskriterien informiert werden (indem sie z. B. die Bewertungskriterien im Internet einsehen können).
Gewichtung	Die Zielsetzung kann dadurch beeinflusst werden, dass den Kandidaten vor dem Test die unterschiedliche Gewichtung der Aufgaben mitgeteilt wird.
Abfolge der Items	In Tests zum Sprechen wird sie normalerweise festgelegt, nicht aber in Tests zum Schreiben.
Zeitliche Begrenzungen	Dies kann sich entweder auf die Vorbereitungszeit beziehen oder auf die Zeit, die zum Lösen der Aufgabe zur Verfügung steht.
Intendierte Tätigkeiten	Allgemeine Beschreibung der sprachlichen Tätigkeiten, die zum Lösen der Aufgaben erforderlich sind. Könnte als redundant angesehen werden, da im folgenden Abschnitt eine ausführliche Liste erstellt werden muss.
Anforderungen: Aufgabe *(Zu beachten: Dies bezieht sich auf die Sprache des INPUTS und auf den ERWARTETEN OUTPUT)*	
Kanal	Der Input kann schriftlich, bildlich (Foto, Kunstwerk usw.), grafisch (Karten, Tabellen usw.) oder mündlich (vom Prüfer oder einer Tonaufnahme) erfolgen. Der Kanal, durch den der Output erfolgt, hängt von der getesteten Fertigkeit ab.
Art des Diskurses	Beinhaltet die Kategorien Genre, rhetorische Anforderung und Darstellungsmuster.
Textlänge	Umfang des Inputs/Outputs.
Schreiber- und Sprecherbeziehungen	Unterschiedliche Beziehungen vorzugeben, kann die Ausführung beeinflussen (z. B. einem bekannten Höhergestellten – wie dem Vorgesetzten – zu antworten, verlangt nicht die gleiche Art von Sprache wie einem Gleichstellten zu antworten).
Art der Information	Abstraktionsgrad. Die Forschung zeigt, dass leichter auf konkrete Themen oder einen konkreten Input zu antworten ist als auf abstrakte Themen oder einen abstrakten Input.
Kenntnis des Themas	Eine bessere Kenntnis des Themas kann zu einer besseren Leistung führen. Dies ist relevant für das Testen aller Fertigkeiten.

Sprache	
Breite des Wortschatzes	Bezieht sich auf die Sprache des Inputs (die normalerweise auf einem Niveau unterhalb des erwarteten Outputs liegt) und auf die Sprache des erwarteten Outputs. Wird durch Bezug auf ein Curriculum oder einen Referenzrahmen wie den Gemeinsamen europäischen Referenzrahmen für Sprachen festgelegt.
Breite der Strukturen	
Breite der Funktionen	
Gesprächspartner	
Sprechtempo	Der erwartete Output sollte muttersprachliche Normen widerspiegeln. Der Input kann dem sprachlichen Niveau der Kandidaten angepasst werden. Dabei besteht jedoch die Gefahr, den natürlichen Rhythmus der Sprache zu verzerren und so eine erhebliche konstrukt-irrelevante Variable einzuführen.
Vielfalt der Akzente	Kann durch die Definition des Konstrukts erforderlich sein (z. B., wenn dort eine Anzahl von Akzenten beschrieben wird) und/oder durch den Kontext (z. B., wenn ein bestimmter Akzent im Unterricht überwiegt).
Bekanntheitsgrad	Es gibt Anzeichen dafür, dass die Leistung der Kandidaten sich verbessert, wenn sie mit einem Freund kommunizieren (obwohl dies möglicherweise kulturabhängig ist).
Anzahl	Betrifft die Merkmale des Kandidaten – es gibt Anzeichen dafür, dass Kandidaten mit unterschiedlichen Persönlichkeitsprofilen sprachlich unterschiedlich handeln, wenn sie mit einer unterschiedlichen Anzahl an Personen kommunizieren.
Geschlecht	Es gibt Anzeichen dafür, dass Kandidaten bessere Leistungen erbringen, wenn sie von einer Frau geprüft werden (auch dies kann kulturabhängig sein), und dass ganz allgemein das Geschlecht des Gesprächspartners das sprachliche Handeln beeinflussen kann.

VALIDITÄT DER ERGEBNISERMITTLUNG	
Kriterien/Bewertungs-vorgaben	Die Kriterien müssen sich auf die Sprachtheorie (das Sprachwissen) beziehen, die im Abschnitt „Theoriebezogene Validität“ und im Abschnitt „Anforderungen: Aufgabe“ genannt wurde. Sie sollten auch die Sprachproduktion widerspiegeln, die von den Testaufgaben gefordert wird.
Bewertungsverfahren	
Beurteilertraining	Für das Training gibt es unterschiedliche Vorgehensweisen. Es gibt Anzeichen dafür, dass ein Training die Bewertungsstrenge, die Konsistenz der Bewertungen und die Fähigkeit verbessern kann, mit den anderen Prüfern/Beurteilern übereinzustimmen.
Standardisierung	Bei jeder Art von Training müssen die Prüfer/Beurteiler die Bestehensgrenze verinnerlichen und dies sollte durch ein Standardisierungsverfahren geprüft werden (als eine Art Test, wenn man so will).
Bewertungsbedingungen	Es sollte versucht werden sicherzustellen, dass das Bewerten/Prüfen unter optimalen Bedingungen stattfindet. Wo dies möglich ist, sollten diese Bedingungen festgelegt werden, damit alle Prüfer/Beurteiler gleichermaßen ihr Bestes geben können.
Adjustierung	Dies bedeutet, dass das Prüfer-/Beurteilerverhalten einer Kontrolle unterworfen wird, um sicherzustellen, dass es konform und konsistent ist.
Statistische Analysen	Statistische Analysen des Beurteilerverhaltens stellen sicher, dass es den Kandidaten nicht zum Nachteil gereicht, wenn Beurteiler oder Prüfer zu streng bzw. zu milde oder nicht konsistent bewerten.
Beurteiler/Prüfer	Wenn wir uns mit den Merkmalen der Lerner befassen (physische und psychologische Merkmale, Erfahrungen), sollten wir uns auch überlegen, was wir über die Beurteiler und Prüfer im Hinblick auf diese Merkmale wissen. In der Forschung gibt es nur wenige systematische Analysen dieser Merkmale mit Blick auf den Prüfer.
Bewertung und Zeugnisvergabe	Die Verfahren, die beschreiben, wie die Endnoten ermittelt und ausgewiesen werden, sollten so explizit wie möglich sein, um sicherzustellen, dass der Test fair ist.

Da der Begriff der kriterienbezogenen Validität, so wie Weir ihn sieht, mit der traditionellen Definition übereinstimmt (vgl. das Kapitel „Kritierienbezogene Validität“), wird dieser Bestandteil des Rahmens hier nicht näher diskutiert.

Der letzte Teil des Rahmens bezieht sich auf die Testwirkungsvalidität. Obwohl nicht völlig klar ist, wie dieser Teil in die übergreifende Validitätsbegründung „passt“, wird die Begründung von Weir kurz skizziert.

Eine eher persönliche Auffassung der Testwirkungsvalidität besagt, dass sie nicht als separates Element existiert, aber stattdessen als ein mehr globaler Aspekt der Testentwicklung angesehen werden sollte, der einen ethischen Ansatz für alle Phasen der Testentwicklung darstellt.

TESTWIRKUNGSVALIDITÄT	
Differentielle Validität	Bezieht sich auf die nachträgliche Analyse der Antwortdaten, um mögliche Fälle von Benachteiligung oder Bevorzugung bestimmter Probandengruppen festzustellen.
Washback im Unterricht oder am Arbeitsplatz	Die Auswirkungen des Tests auf das Lernen und den Unterricht. Tests wurden zum Teil eingesetzt, um den Unterricht zu beeinflussen, obwohl sich gezeigt hat, dass dieser Einfluss an der Oberfläche bleibt, wenn die Lehrer nicht am Revisionsprozess beteiligt werden. Obgleich es eine Menge Beweise für Washback gibt, liegt bis heute keine Beschreibung oder Definition dessen, was Washback genau ist, vor (siehe aber Cheng, Watanabe & Curtis (2004), in dem eine Reihe von Untersuchungen genannt werden, die das Verständnis dieses Begriffs vorangebracht haben).
Auswirkung auf den Einzelnen innerhalb der Gesellschaft	Die Auswirkungen zu untersuchen, die ein Test auf die Gesellschaft hat, ist vermutlich das schwierigste Unterfangen von allen und das vermutlich am häufigsten vernachlässigte, da eine solche Untersuchung über die unmittelbar vom Test Betroffenen hinausgeht.

Weirs Rahmen wurde dazu verwendet, existierende Prüfungen zu beschreiben (siehe Weir und Shaw sowie O'Sullivan und Green, erscheinen demnächst). Sie dienten als Grundlage für den Nachweis der Spezifizität von Sprache für Tests mit einem spezifischen Verwendungszweck (siehe O'Sullivan, 2005) und als Grundlage für die Erstellung von detaillierten Testbeschreibungen (siehe QUALSPELL, 2005 und EXAVER, 2005). Sie wurden auch als Grundlage für die Entwicklung von Prüfungen in anderen Fächern verwendet (z. B. für den Einstufungstest für Mathematik in den Vereinigten Arabischen Emiraten). Der Einfluss, den diese Rahmenvorgaben besonders in Europa haben, wächst, was zum Teil daran liegt, dass sie verhältnismäßig einfach auf unterschiedliche Prüfungen und unterschiedliche Kontexte anzuwenden sind.

6. ÜBUNGEN

1. Lesen Sie die folgenden Zitate, und beantworten oder diskutieren Sie die folgenden Fragen.

Inwieweit stimmen Sie den Aussagen zu?
Was sind die Auswirkungen für die Testerstellung?

1.
Die wichtigste Funktion von Sprache ist die Interaktion und die Kommunikation.
(Richards und Rodgers, 1988: 70)

2.
Das hauptsächliche Problem beim Lernen einer Fremdsprache besteht darin, die Struktur dieser Sprache zu meistern, und dieses Problem erfordert, dass man ihm eine nahezu uneingeschränkte Aufmerksamkeit schenkt.
(Roberts, 1982: 99)

3.
Sprache ist ein System von Gewohnheiten. Das Erlernen einer Fremdsprache ist im Grunde ein Prozess der Ausbildung automatisierter Gewohnheiten.
(Richards und Rodgers, 1988: 51, über die audio-linguale Methode)

4.
Kommunikation beinhaltet Freiheit und Unvorhersehbarkeit.
(Xiaoju, 1990: 61)

5.
Unterrichtsmaterialien sollten immer so authentisch wie möglich sein.
(Wright, 1987: 76)

2. Ziehen Sie als Beispiel zwei Fremdsprachentests heran, die Sie kennen.

Versuchen Sie, jeden Test in Beziehung zu den Auffassungen von Sprache oder zu den Modellen der Sprachkompetenz zu setzen, die in diesem Modul vorgestellt wurden.

ANHANG A

Literaturempfehlungen

Die folgende kurze Liste enthält einige der wichtigsten Bücher, auf die sich dieses Modul stützt. Eine ausführlichere Bibliographie findet sich in **Anhang B**.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

EXAVER. (2005). Project Website. http://www.uv.mx/exaver/index.html. Webseite abgerufen im April 2005.

Heaton, J. B. (1975). *Writing English Language Tests*. London: Longman.

Lado, R. (1961). *Language Testing*. London: Longman.

McNamara, T. (2003). *Validity and Reliability in the Senior School Curriculum: new takes on old questions*. Invited Presentation, Australian Curriculum, Assessment & Certification Authorities (ACACA), National Conference, Adelaide, 31. Juli 2003.

Mislevy, R. (2003). *On the Structure of Educational Assessments*. CSE Technical Report 597. University of California, Los Angeles: Centre for the Study of Evaluation.

Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2005). *Issues in Business English Testing: the BEC revision project*. Cambridge: Cambridge University Press.

O'Sullivan, B. & Green, A. (erscheint demnächst). *Examining Speaking*. Cambridge: Cambridge University Press.

QALSPELL. (2005). *Project Handbook.*

van Ek, J. A. & Trim, J. L. M. (1991). *Threshold Level 1990*. Strasbourg: Council of Europe.

Weir, C. & Shaw, S. (erscheint demnächst). *Examining Writing*. Cambridge: Cambridge. University Press
Weir, C. (2005). *Language Testing and Validation: an evidence-based approach*. Oxford: Palgrave.

Wilkins, D. (1976). *Notional Syllabuses*. Oxford: Oxford University Press.

ANHANG B

Bibliographie

Alderson, J. C. (1981). Reaction to the Morrow paper (3). In: Alderson, J. C & Hughes, A. (eds.). (1981). *Issues in Language Testing*. ELT Documents 111. London: The British Council.

Alderson, J. C. & Hughes, A. (1981). *Issues in Language Testing*. ELT Documents 111. London: The British Council.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Blalock, H. M. (1980). *Sociological Theory and Research*. New York: Macmillan.

Bolton, S. (1985). *Die Gütebestimmung kommunikativer Tests*. Tübingen.

Brumfit, C. (1980). *Problems and Principles in English Language Teaching*. Oxford: Pergamon Press.

Canale, M. & Swain, M. (1981). A theoretical framework for communicative competence. In: Palmer, A. S., Groot, P. G. and Trosper, S. A. (eds.). *The construct validation of tests of communicative competence.* Washington, D. C., TESOL, 31-36.

Carroll, J. B. (1968). 'The psychology of language testing'. In: Davies, A. (1968): 46-69.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. New York: Harper and Row.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers in Bilingualism*. 19: 97-205.

Davies, A. (ed.) (1968). *Language Testing Symposium. A Psycholinguistic Perspective*. London: Oxford University Press.

Finocchiaro, M. & Brumfit, C. (1983). *The Functional-Notional Approach: From Theory to Practice*. Oxford: Oxford University Press.

Gulliksen, H. (1987). *A Theory of Mental Tests*. Hillsdale, New Jersey: Lawrence Erlbaum.

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill.

Heaton, J. B. (1975). *Writing English Language Tests*, London: Longman.

Hymes, D. (1972). On communicative competence. In: Pride, J. B. & Holmes, J. (eds.). *Sociolinguistics*. Harmondsworth: Penguin.

Lado, R. (1961). *Language Testing.* London: Longman.

Maley, A. (1986). A rose is a rose, or is it?: Can communicative competence be taught? In: Brumfit, C. (ed.). (1986). *The Practice of Communicative Teaching*. ELT Documents 124. Oxford: Pergamon Press.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In: Brumfit, C. J. & Johnson, K. (eds.). The Communicative Approach to Language Teaching. Oxford: Oxford University Press.

Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Nitko, A. J. (1983). *Educational Tests and Measurement, An Introduction*. New York: Harcourt, Brace, Jovanovich.

Oller, J. W. (1991). Foreign Language Testing, Part 1: Its Breadth, *ADFL Bulletin*, 22: 33-38.

Richards, J. C. (1990). Communicative needs in foreign language teaching. In: Bolitho, R. & Rossner, R. (eds.) (1990). *Currents of Change in English Language Teaching*. Oxford: Oxford University Press: 48-58.

Richards, J. C. & Rodgers, T. (1988). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Roberts, J. T. (1982). Recent developments in ELT. In: Kinsella, V. (ed.). (1982). *Surveys 2*. Cambridge, Cambridge University Press.

Spolsky, B. (1975). Language testing: Art or science? Paper presentation. Stuttgart: fourth AILA International Congress.

Swan, M. (1990). A critical look at the communicative approach. In: Bolitho, R. & Rossner, R. (eds.). (1990). *Currents of Change in English Language Teaching*. Oxford: Oxford University Press: 73-98.

Valette, R. M. (1967). *Modern Language Testing: A Handbook*. New York: Harcourt, Brace and World.

van Ek, J. A. (1975). *Threshold Level*. Strasbourg: Council of Europe.

van Ek, J. A. & Trim, J. L. M. (1991). *Threshold Level 1990*. Strasbourg: Council of Europe.

Weir, C. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Wilkins, D. (1976). *Notional Syllabuses*. Oxford: Oxford University Press.

Wright, T. (1987). *Roles of Teachers and Learners*. Oxford: Oxford University Press.

Xiaoju, L. (1990). In defence of the communicative approach. In: Bolitho, R. & Rossner, R. (eds.). *Currents of Change in English Language Teaching*. Oxford: Oxford University Press: 59-72.