Testen und Prüfen in der Fremdsprache: Was macht eine gute Sprachprüfung aus?

Foreign Language Testing and Evaluation: What defines a good test?

Jan Stevener[1]

Abstract

This paper aims to outline essential concepts and criteria in order to assess the quality and aptitude of language tests for any given specific goal. For this purpose, vital terms related to language-assessment and evaluation are explained and discussed. Subsequently, this paper describes six points which enable test users and authors to assess the aptitude of a test for a specific goal. Among these points are clarity about the research object, the test construct, the properties that researchers assign to it and how these can be measured, as well as how adequate the research design is to evaluate the research object. Moreover, the application of well-established quality criteria like validity, objectivity, reliability, etc. facilitates the evaluation of language tests. It is shown that quality criteria interact with each other and that test users and authors have to weigh them according to the purpose of the test.

Keywords: Language Testing; Evaluation; Assessment; Data Collection; TestDaF; DSH; Quality Criteria

[1]Assistant Lecturer from the Humanities and Language Department of Mahidol University International College (MUIC), Humanities and Language Division, Mahidol University, Salaya, Nakhon Pathom. Email: janstevener@yahoo.de

Abstrakt

Der vorliegende Beitrag stellt zentrale Konzepte und Kriterien zur Einschätzung der Eignung und Qualität einer Sprachprüfung für spezifische Zielsetzungen vor. Zu diesem Zweck werden zuerst zentrale Begriffe der Leistungsmessung vorgestellt und diskutiert. Anschließend werden sechs zentrale Punkte erläutert, die Prüfungsanwender oder –hersteller in die Lage versetzen sollen, die Eignung einer Prüfung für einen spezifischen Zweck einzuschätzen. Dazu zählen Klarheit über den Untersuchungsgegenstand und das Testkonstrukt sowie die damit verbunden Eigenschaften, in welchen Maßen sich diese messen lassen und wie geeignet die Operationalsierung zur Erfassung des Untersuchungsgegenstandes ist. Des Weiteren soll mit Hilfe etablierter Gütekriterien wie Validität, Objektivität, Reliabilität etc. eine Einschätzung der Qualität einer Prüfung ermöglicht und gezeigt werden, dass Gütekriterien sich wechselseitig beeinflussen und der Testersteller oder –anwender diese je nach Zielsetzung unterschiedlich gewichten kann.

Schlüsselwörter: Testen, Prüfen, Fremdsprachenprüfungen, Datenerhebung, TestDaF, DSH, Gütekriterien

1. Einleitung

Lehrende im Fremdsprachenunterricht verbringen beträchtliche Zeit damit, Lernende zu evaluieren. Dies geschieht in unterschiedlichsten Formen und reicht vom spontanen Test zur Lernstandskontrolle bis zu komplexen formellen Abschlussprüfungen. Es kann um die Evaluation einzelner Teilkompetenzen gehen oder um die Erfassung kombinierter Fertigkeiten mit Hilfe von Testbatterien. Für die Lernenden können die Ergebnisse solcher Evaluationen erhebliche persönliche Konsequenzen haben. Das Ergebnis kann über eine Versetzung in eine höhere Schulklasse entscheiden, über die Aufnahme eines Hochschulstudiums in Deutschland, sogar über die Erteilung eines Visums zum Zwecke der Eheschliessung oder die Erteilung einer Staatsbürgerschaft.

In vielen Programmen an Universitäten in Thailand ist die Note für einen Kurs identisch mit den Ergebnissen der Zwischenprüfung und der Abschlussprüfung. Letzendlich entscheiden damit Zwischen- und Abschlussprüfungen über die Note im Bachelor oder Master und damit auch über die Optionen, die ein Absolvent nach dem Studium hat. Es muss daher ein Merkmal des Testens und Prüfens sein, dass das verwendete Testverfahren kritisch reflektiert und die Grenzen des möglichen Erkenntnisgewinns deutlich werden, denn „eine voreilige Diagnose von sprachlichen Fertigkeiten ohne die wissenschaftliche Fundierung solcher Tests kann verhängnisvoll werden“ (Roche, 2013, p. 108). Bevor jedoch Kriterien beschrieben werden, mit denen die Qualität einer Prüfung oder eines Tests erfasst werden kann, sollen die zentralen Begriffe näher beschrieben werden.

2. Zentrale Begriffe: Test und Prüfung

In der Praxis, aber auch der Fachliteratur, werden die Begriffe *Prüfen/Prüfung* und *Testen/Test* nicht einheitlich verwendet. Vor allem in älterer Fachliteratur findet man das Bemühen, die beiden Begriffe durch das Merkmal *formell* (Prüfung) und *informell*

(Test) zu unterscheiden. Demzufolge beziehen sich informelle Tests nur auf eine bestimmte Gruppe von Lernenden und einen begrenzten Lehrstoff. Sie werden von den Lehrenden entworfen und oft ad-hoc ohne besondere Vorbereitung durchgeführt. Testerstellung und Bewertung orientieren sich nicht an offiziellen Kriterien (Bolton, 1996, p. 6). Prüfungen werden hingegen als formell eingestuft. Sie beanspruchen, unabhängig von der Bezugsgruppe, den eingesetzten Lernmedien und –methoden Auskunft über das sprachliche Können zu geben. Formelle Prüfungen orientieren sich an festgelegten Kriterien und können daher objektive Aussagen in Bezug auf diese Kriterien liefern (Albers & Bolton, 1995, p. 14). Diese Unterscheidung wird jedoch nicht strikt befolgt. Während beispielsweise die DSH (Deutsche Sprachprüfung für den Hochschulzugang) noch die Prüfung im Namen trägt, ist der jüngere TestDaF, der den Test im Namen trägt, nach den obengenannten Bestimmungen eindeutig eine formelle Prüfung. Ferner scheint der Bezug zur nicht-deutschen Fachliteratur, in der vom *test* gesprochen wird, einigen Einfluss zu haben. Unter anderem sind PET (Preliminary English Test), TCF (Test de Connaissance du Francais) oder TOEFL (Test of English as a Foreign Language) zu nennen. Einer der größten Anbieter für Sprachzertifikate nach dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (im Folgenden als GER bezeichnet) des Europarates (2001), die TELC (früher: Weiterbildungs-Testsysteme GmbH), bezeichnet die angebotenen Zertifikate auf der Homepage als *language tests* (2018). Besonders wenn es um die Vergabe von Zertifikaten geht, wird zwar eher von Prüfungen gesprochen, jedoch verwenden Anbieter solcher Prüfungen auch zunehmend Bezeichnungen, die neutral sind: Fit in Deutsch, Start Deutsch 1, Kleines/Großes Deutsches Sprachdiplom, Zertifikat Deutsch für den Beruf etc. Auch neuere Fachliteratur verzichtet eher auf eine strenge Trennung beider Begriffe (vgl. Grotjahn & Kleppin,

2015). Für diesen Beitrag habe ich mich daher entschieden, vor allem den Begriff Prüfung und Prüfen zu verwenden, da die für formelle Prüfungen geforderte Kriterienorientierung dem Ziel dient, die Qualität einer Sprachprüfung besser bestimmen zu können und auch bei informellen Tests beachtet werden sollte. Gemeinsames Merkmal von Tests und Prüfungen ist es, dass Prüflinge durch eine Aufgabenstellung zu bestimmten sprachlichen Handlungen oder Reaktionen gebracht werden sollen. Es handelt sich um den Einsatz von „theoretisch und empirisch fundierten Verfahren zur kontrollierten Auslösung von diagnostisch relevantem Verhalten durch standardisierte Reize – mit dem Ziel eines Rückschlusses auf sprachliche Kompetenzen“ (Grotjahn, 2013, p. 211). Bei alternativen Formen der Leistungsmessung, z.B. durch Unterrichtsbeobachtung oder Selbstevaluation, verwendet man hingegen die Begriffe *Evaluation* oder im Englischen den Begriff *Assessment.* Ferner wird zwischen summativer und formativer Evaluation unterschieden. Summative Evaluationen sind punktuell und produkt-/ergebnisorientiert, beispielsweise als Noten in Zeugnissen, während formative Evaluationen kontinuierlich und prozessorientiert im Unterricht integriert sind. Sie dienen dazu, den Unterricht zu optimieren und die Fertigkeiten der Lernenden weiterzuentwickeln (Grotjahn & Kleppin, 2015, p. 36). Im Grunde ist somit schon die Beobachtung im Unterricht, dass es bei den Lernenden bestimmte Defizite gibt, eine formative Evaluation.

Verfahren, die auf Grund von Zahlenwerten Rückschlüsse auf sprachliche Fertigkeiten liefern, werden oft als Messverfahren bezeichnet. Letzlich liefern die meisten Tests und Prüfungen Leistungsbeurteilungen auf Grund von Zahlenwerten und müssen daher als Messverfahren gelten. So beschreibt Kecker Sprachprüfungen weiterführend als „Messverfahren, die in möglichst sytematischer und objektiver Weise Aussagen über die

Sprachfähigkeit von Personen treffen oder über ihre Fähigkeit, bestimmte sprachliche Aufgaben zu bewältigen“ (Kecker, 2011, p. 26). Für eine Leistungsbeurteilung müssen nicht direkt beobachtbare theoretische Konstrukte, z.B. kommunikative Kompetenz oder Hörverstehenskompetenz, in quantifizierbare und zählbare Ergebnisse überführt werden. Bei der Überführung muss mit großer Achtsamkeit vorgegangen werden, denn es sollte erkennbar sein, dass Überlegungen dazu angestellt worden sind, welcher Erkenntnisgewinn mit einer bestimmten Prüfung überhaupt möglich ist und wie zuverlässig und objektiv gemessen wird. Im Folgenden möchte ich daher Kriterien vorstellen, die Voraussetzungen, Möglichkeiten und Grenzen von Prüfungen eruieren und einen kritischen und reflektierten Umgang mit Prüfungen ermöglichen.

3. Kriterien und deren testtheoretische Funktionen

Die Kriterien erfüllen eine wichtige Funktion bei der Beantwortung der Frage, was untersucht werden soll und kann. Sprachprüfungen beanspruchen zumeist, bestimmte Kompetenzbereiche zu überprüfen (Hörverstehen, Leseverstehen, Sprechen, Schreiben, Hör-Seh-Verstehen, Sprachmittlung etc.). Kompetenzen können verschiedene Teilkompetenzen voraussetzen (Beispiele finden sich u.a. in Hallet, 2008 und Traoré, 2016). Die explizite Benennung von Kompetenzen als Untersuchungsgegenstand hat vor allem damit zu tun, dass moderner Fremdsprachenunterricht als kompetenzorientiert oder handlungsorientiert beschrieben wird. Einen nicht zu unterschätzenden Einfluss auf diese Entwicklung hatte wohl auch der Gemeinsame Europäische Referenzrahmen (GER), der die sprachlichen Niveaustufen A1 bis C2 als Kompetenzniveaus mit Hilfe von *Kann*-Beschreibungen definiert. Seither verorten

Lehrmaterialien und Kurse ihre Niveaustufen gemäß den Kompetenzniveaus im GER, selbst methodisch hochentwickelte Prüfungen wie der TestDaF, der gut fundierte eigene Niveaustufen (TDN 3 - TDN 5) aufweisen kann. Ziel des Fremdsprachenunterichts ist die Vermittlung von bestimmten Kompetenzen, und der Erfolg der Sprachvermittlung kann durch eine Erhebung dieser Kompetenzen überprüft werden. Klare Aussagen darüber, was die Lernenden schon können sollen, bilden eine sehr wichtige Grundlage für die lernzielgerechte Überprüfung dieser (Teil-) Kompetenzen. Dabei sollte nicht vergessen werden, dass man Kompetenzen nicht direkt erfassen kann. Erfasst werden kann Kompetenz nur über beobachtbares Verhalten. Die Kompetenz an sich ist nur ein theoretisches Konstrukt, dem vom Testentwickler bestimmte Ausprägungen als beobachtbares Verhalten zugeschrieben werden. Kompetenzen, die in einer Prüfung gemessen werden sollen, werden daher als *Testkonstrukte* bezeichnet.

Die Explizierung des Testkonstrukts schafft Klarheit über den zu prüfenden Phänomenbereich. Entwickler sollten in Zusammenhang mit der Operationalisierung transparent machen, welche Merkmale sie diesem Konstrukt theoretisch zuschreiben und wie sich diese empirisch messen lassen. Der Zusammenhang von Theorie und Empirie soll so offengelegt werden. Damit wird Testanwendern die Möglichkeit gegeben, oft nur implizite theoretische Grundannahmen zu hinterfragen und die Nachvollziehbarkeit und Transparenz der Prüfung erhöht. Grotjahn & Kleppin (2015, p. 87) verweisen auf die genaue Beschreibung des Testkonstruktes auch als Voraussetzung für die Auswahl geeigneter Aufgabenformate. Ferner sollen die klassischen Gütekriterien Validität, Reliabilität und Objektivität auf den Test angewendet werden, denn erst mit Hilfe der Gütekriterien wird deutlich, welcher Erkenntnisgewinn mit der eingesetzten Sprachprüfung überhaupt zu erzielen ist. Im

Rahmen dieses Artikels können jedoch nur in sehr beschränktem Umfang Beispiele gezeigt und Kriterien praktisch angewendet werden.

Die nun folgenden Punkte 3.1 bis 3.6 sollen daher vor allem helfen, eine größere testtheoretische Reflektiertheit und Transparenz bei Lehrenden, die Tests entwicklen und anwenden möchten, zu erreichen:

3.1 Kurzbeschreibung des Verfahrens

Eine Kurzbeschreibung des Verfahrens ermöglicht potenziellen Anwendern, ein erste Einschätzung bezüglich der Eignung des Verfahrens für eine bestimmte Fragestellung als auch eine Einschätzung zur Durchführbarkeit vorzunehmen. Dazu gehört es, einerseits kurz den Gegenstand der Prüfung zu beschreiben und andererseits die Form der Prüfung und seiner Aufgaben zu klassifizieren. Soll das Verfahren beispielsweise den Mündlichen Ausdruck (Gegenstand) eines Probanden erheben, so kann dies mit einem Interview (Form) erfolgen. Die Aufgaben wären als offene, halboffene oder geschlossene Fragen zu klassifizieren. Eine Prüfung kann sich dabei aus verschiedensten Einzelprüfungen zusammensetzen, insbesondere bei Einstufungsprüfungen, Eignungsprüfungen oder Zulassungsprüfungen (DSH, TestDaF, TOEFL etc). Es ist zu beachten, dass die Einzelprüfungen einzeln beschrieben werden. Ein Vorbild ist z.B. der TestDaF, bei dem den Prüflingen vor dem Prüfungsteil schriftlich und auditiv der Gegenstand benannt und kurz beschrieben wird, welche Aufgaben zu erwarten sind: „Im Prüfungsteil Mündlicher Ausdruck sollen Sie zeigen, wie gut Sie Deutsch sprechen. Dieser Teil besteht aus insgesamt 7 Aufgaben, in denen Ihnen unterschiedliche Situationen aus dem Universitätsleben vorgestellt werden. Sie sollen sich zum Beispiel informieren, Auskunft geben oder Ihre Meinung sagen. Jede Aufgabe besteht aus zwei Teilen: Im ersten Teil wird die Situation beschrieben, in der Sie sich befinden, und es wird gesagt, was Sie tun sollen ... im zweiten Teil der

Aufgabe spricht Ihr Gesprächspartner oder Ihre Gesprächspartnerin, danach sollen Sie sprechen“. Das Beispiel stammt aus der Musterprüfung 1 (TestDaF-Institut, 2005, p. 41).

3.2 Gegenstand der Prüfung

Dieser Punkt erfordert eine Benennung des Untersuchungsgegenstandes aus theoretischer Sicht. Erst wenn Klarheit über den zu untersuchenden Phänomenbereich herrscht, kann entschieden werden, ob eine Prüfung in der Lage ist, über diesen Gegenstand Auskunft zu geben oder ob die Prüfung modifiziert oder sogar um weitere Verfahren ergänzt werden muss. Zur Verdeutlichung können hier die beiden bekanntesten Sprachprüfungen für die Zulassung zum Studium in Deutschland angeführt werden, die DSH und der TestDaF. Beide beanspruchen, die sprachliche Fähigkeit von ausländischen Studienbewerbern zum Studium an einer deutschen Hochschule als Untersuchungsgegenstand zu erheben. Beide Prüfungen gelten als Testbatterien, da sie verschiedene Einzelprüfungen umfassen: Mündlicher und schriftlicher Ausdruck, Hör- und Leseverstehen. Ein wichtiger Unterschied ist jedoch, dass TestDaF auf die gesonderte Erhebung des Phänomenbereichs „Grammatik/Strukturen“ verzichtet. Während man einwenden kann, dass wissenschafts-sprachliche Strukturen mit bestimmten grammatischen Formen einhergehen und daher ein Prüfungsteil Grammatik nötig sei, kann ebenso argumentiert werden, dass solche wissenschaftssprachlichen Strukturen bereits bei den anderen 4 Prüfungsteilen erhoben werden. Eine Einschätzung, welche Prüfung den genannten Gegenstand am besten erhebt, ist jedoch nur mit einer transparenten und expliziten Benennung des Gegenstandes möglich.

Sprachprüfungen erfassen in aller Regel Kompetenzen als Untersuchungsgenstand. Wie bereits erwähnt lassen sich diese Kompetenzen nicht direkt messen. Kompetenzen, die beobachtbarem Verhalten zugrunde liegen, werden häufig als

Testkonstrukte bezeichnet (Grotjahn & Kleppin, 2015). Da jedoch das Konstrukt nicht umittelbar messbar ist, kann nur das sprachliche Handeln und Verhalten beobachtet werden; daher spricht man auch von Performanztests. Anschließend sollten Rückschlüsse auf die zugrundeliegende Kompetenz möglich sein. Ein Beispiel ist die Prüfung der Sprechfähigkeit mittels eines Rollenspiels, wie in der Prüfung Goethe Zertifikat B1 des Goethe Instituts oder simulierte Telefongespräche wie im Prüfungsteil Mündlicher Ausdruck des TestDaF. Ferner ist zu beachten, dass beim TestDaF oder der DSH sprachliche Kompetenzen für einen bestimmten Bezugsbereich, das Studium, geprüft werden sollen. Neben einer exakten Benennung der Kompetenzen setzt dies auch eine Definition des Bezugsbereichs voraus (Chapelle et al, 2010, p. 8).

Insbesondere bei der Überprüfung rezeptiver Kompetenzen (Lese- oder Hörverstehen) kann nur sehr indirekt durch Interferenz auf Kompetenzen geschlossen werden, beispielsweise durch das Ankreuzen in einer Mehrfachauswahlaufgabe. Bei solchen Tests wird häufiger von Kompetenztests gesprochen (Grotjahn, 2013, p. 213). Werden Prüfungen für den Einsatz als Abschluss- oder Lernfortschrittstest konzipiert, so kann das Testkonstrukt passgenau auf die Lernziele des vorangegangenen Unterrichts abgestimmt werden. Prüfungen, die das Erreichen der Niveaustufen A1 – C2 überprüfen, können sich mit einer gewissen Vorsicht auf die im GER (Europarat, 2001) definierten Niveaustufen beziehen. Der GER bietet zwar Beschreibungen der jeweiligen Kompetenzstufen, jedoch warnen Fulcher (2004, 2010) und Milanovic (2009, p. 3) vor einer Überschätzung der Möglichkeiten des GER. Das Konstrukt einer Prüfung werde nicht vom GER bestimmt, sondern ein Testkonstrukt in einem spezifischen Kontext kann in Hinblick auf seine Übereinstimmung mit dem GER überprüft werden. Ferner wurde kritisiert, dass die in den Skalen verwendeten Begriffe, wie „vertraut“, „einfach“ oder „komplex“ nicht

erläutert und Schwierigkeitsmerkmale nicht kohärent verwendet werden (Alderson & Hutha, 2005).

3.3 Operationalisierung

Die Frage, welcher Gegenstand wie gemessen wird, ist zentral für Entscheidungen zur Operationalisierung der Forschungsfrage: „Die Operationalisierung eines theoretischen Begriffs besteht aus der Angabe einer Anweisung, wie Objekten mit Eigenschaften (Merkmalen), die der theoretische Begriff bezeichnet, beobachtbare Sachverhalte zugeordnet werden können“ (Schnell, Esser, & Hill, 1995). Es ist beispielsweise fraglich, ob Mehrfachauswahlaufgaben (*multiple choice* Aufgaben) tatsächlich sprachliche Kompetenzen erfassen können, oder ob nicht eher die Vertrautheit der Prüflinge mit eben diesem Testformat erfasst wird (Perlemann-Balme, 2001). Um die Eignung einer Prüfung einschätzen zu können, müssen die ihr zugrundeliegenden theoretischen Annahmen offenbart werden.

Wenn beispielsweise erhoben werden soll, wieweit die Prüflinge bereits ihre Sprachverwendung automatisiert haben, so muss deutlich werden, welche Merkmale diesem Untersuchungsgegenstand theoretisch zugeschrieben werden. Die Operationalisierung muss sich also auf die Frage beziehen, wie sich diese Eigenschaften messen lassen. Prüfungen, die Niveaustufen nach dem GER erfassen möchten, können auf die dortigen *Kann-Beschreibungen* zurückgreifen. Diese sind jedoch recht abstrakt und oft frei interpretierbar. Daher empfiehlt sich für die Operationalisierung der Einsatz von „Profile Deutsch“ (Glaboniat et al, 2005). Profile Deutsch beruht auf den Kann-Beschreibungen des GER, konkretisiert diese jedoch für die Praxis und erweitert sie durch Beispiele.

Das folgende Beispiel ist aus einer *High-Stakes* Prüfung für den Hochschulzugang (Professional Aptitude Test, PAT 7.2, 2011). Im Prüfungsteil *Grammatik* gibt es die folgende Mehrfachauswahlaufgabe mit der Arbeitsanweisung „Wählen Sie die beste Antwort!“:

Was, du _____ heiraten?

Das _____ nicht wahr sein!

a) möchtest/muss

b) möchtest/soll

c) willst/darf

d) willst/mag

Neben der bereits genannten Kritik an der Eignung vom *multiple choice* Aufgaben zur Erfassung vom Kompetenzen ist hier zu bemerken, dass keine Grammatik erfasst wird, weil die Stellung der Verben im Satz vorgegeben ist und die Verben flektiert sind. Bei Fragestellungen der Operationalisierung kann man also auch deduktiv fragen, welche Eigenschaften für die Lösung dieser Aufgabe nötig sind. „Das darf nicht wahr sein“ kann als feststehende Redewendung zum Ausdruck von Überraschung betrachtet werden, denn das Modalverb dürfen wird eben nicht in der modalen Bedeutung „erlauben“ verwendet. Insofern geht es weder um Semantik noch im Speziellen um die Bedeutung der Modalverben, sondern vielmehr um Vertrautheit mit Redewendungen. Da es sich um eine feststehende Redewendung handelt, ist ferner auch die Arbeitsanweisung irreführend, da nur eine Antwort korrekt ist. Ferner kann an Hand von Profile Deutsch ermittelt werden, dass das Verstehen durch schriftliche Rezeption von oft gebrauchten Wendungen auf dem Niveau B1 verortet wird (p. 105), während die Beherrschung der Modalverben den Niveaustufen A1 (wollen, müssen, mögen) und A2 (dürfen, sollen) zugeschrieben wird (Glaboniat et al, 2005).

Die Offenlegung der Operationalisierung macht es auch möglich, Faktoren, die das Ergebnis verfälschen, leicher zu erkennen. Es kann sich dabei um Messfehler handeln, beipielsweise wenn der Einsatz von Wörterbüchern oder Smartphones während der Prüfung nicht geregelt ist und so manche Prüflinge bessere Ergebnisse erzielen können als andere. Andererseits kann es sich um Störfaktoren handeln, die oft nicht beeinflussbar sind. Dies kann z.B. die fehlende Vetrautheit mit einem bestimmtem Testformat sein. Man denke an den TestDaF: Im Prüfungsteil

mündlicher Ausdruck muss der Prüfling auf eine Frage monologisch die Antwort in das Aufnahmegerät sprechen, zudem gibt es auch eine strikte Zeitvorgabe für die notwendigen Handlungen *Überlegen* und *Sprechen*. Ist der Prüfling nicht auf dieses Format vorbereitet, ist eine schlechtere Performanz als in einer natürlichen Situation erwartbar. Andere Störfaktoren können unterschiedliche Belastbarkeit der Teilnehmer, unterschiedliche Lösungsstrategien etc. sein, auf Seiten der Prüfenden Vorlesende mit dialektaler Färbung (DSH) usw., ebenso zufällige Störfaktoren in der Durchführung: Störgeräusche aus Nebenräumen, ungünstige Sitzplätze bei Hörverstehensaufgaben, Ausfall technischer Geräte. All diese Faktoren beeinflussen die Zuverlässigkeit, mit der gemessen wird. Da diese nicht beeinflussbar sind, sollten sie zumindest dokumentiert werden.

3.4 Standardisierung

Eine zentrale Forderung für eine gute Prüfung ist eine hohe Standardisierung. Ziel der Standardisierung ist eine Vereinheitlichung der Prüfung, bei der die Bedingungen, unter denen die Prüfung abgelegt wird, so vergleichbar wie möglich gemacht werden. So muss z.B. festgelegt werden, welche Texte und welche Aufgabenformate verwendet werden, welche sprachliche Handlungen der Prüflinge als erfolgreich zu sehen sind und wie diese Reaktionen zu bewerten sind. Die Standardisierung ist insbesondere unerlässlich für formelle Prüfungen wie den TestDaF, die DSH oder TOEFL, denn diese Prüfungen werden mehrfach pro Jahr angeboten. Dabei muss gewährleistet bleiben, dass beispielsweise ein Ergebnis von TDN 4 im Leseverstehen des TestDaf aus einem Prüfungsdurchgang vergleichbar bleibt mit einem TDN 4 aus einem Prüfungsdurchgang im folgenden Jahr, bei dem andere spezifische Items verwendet wurden. Die Standardisierung im Leseverstehen des TestDaF erfolgt daher an Hand der folgenden Kriterien: Dauer, Anzahl der Items, Textlänge, Itemtyp (Zuordnung, multiple choice, ja/nein/Text sagt nichts dazu), Diskursart, Aspekte des Leseverstehens

(Globalverstehen, Einzelheiten verstehen, Inferenzen ableiten) und Zielsetzung des Leseverstehens (zur Orientierung lesen, Information und Argumente verstehen, Gedankengang in einem Text verstehen, implizite Bedeutungen verstehen) (Kecker, 2011, p. 141). Auf diese Weise bleibt trotz unterschiedlicher Items in der jeweiligen Prüfung das getestete Leseverstehen vergleichbar. Im Vergleich ist die DSH deutlich schwächer standardisiert als TestDaF, da es zwar eine Rahmenordnung für die Durchführung gibt, jedoch die Auslegung dieser Regelungen Ermessensspielräume lässt, die von den Hochschulen unterschiedlich ausgelegt werden; so kann der Hörverstehenstext an einer Hochschule vom Band kommen, an einer anderen Hochschule aber von einem Prüfer vorgelesen werden, wobei Variablen wie Aussprache, Lesegeschwindigkeit, dialektale Färbung usw. nicht standardisert sind und je nach Vorlesendem anders ausfallen werden. So bezeichnet Hallet die Standardisierung als „eine zentrale Voraussetzung für eine zufriedenstellende Objektivität“ (2013, p. 212) der Prüfung. Weiterführendes dazu findet sich unter 3.2.

3.5 Maße

Die Offenlegung der Maße soll Klarheit darüber schaffen, welche Ausprägung die beobachtbare Eigenschaft, die dem Gegenstand der Prüfung zugeschrieben wird, annehmen kann. Die verwendeten Maße dienen damit der Quantifizierung der beobachteten Sachverhalte. Wenn beispielsweise erfasst werden soll, inwieweit ein Prüfling die mündliche Sprachverwendung automatisiert hat, so ist „Automatisierung“ ein theoretisches Konstrukt, dem u.a. die beobachtbaren Eigenschaften „Schnelligkeit“ und „Mühelosigkeit“ zugeschrieben werden (Stevener, 2003). Diese Eigenschaften müssen zählbar werden. Als Maße für Schnelligkeit werden vor allem temporale Variablen untersucht. Daher kann z.B. die Artikulationsrate (die Anzahl von Silben pro Minute, Pausen herausgerechnet), die Anzahl

gefüllter und ungefüllter Pausen, die sog. *speech rate* (das Verhältnis von Silben zur Gesamtdauer der Äußerung inklusive Pausen) etc. verwendet werden.

Sprachprüfungen wollen bestimmte Kompetenzen erfassen. Dabei muss deutlich werden, auf welche Eigenschaften die Prüfung zielt und welche Maße gewählt wurden. Wird z.B. die Schreibkompetenz untersucht, so dürften viele Lehrende annehmen, dass die beobachtbare Eigenschaft „Korrektheit" an Hand der Anzahl von Fehlern gemessen werden kann und so ein Rückschluss auf Schreibkompetenz möglich ist. Diese Fehler können unterschiedlich gewichtet werden. Darüber hinaus können der Schreibkompetenz aber weitere Eigenschaften zugeschrieben werden, z.B. Komplexität, Angemessenheit, Kohärenz, etc. Die Offenlegung der verwendeten Maße ermöglicht eine genauere Einschätzung, was gemessen wird und in welchem Verhältnis diese Maße zum Gegenstand der Prüfung stehen. Spolsky (2000, p. 539) schreibt treffend: „One of the easiest things to do, it has been suggested, is to develop a new kind of test – what is hard to know is to know what an existing test really measures". Die Maße quantifizieren die beobachteten Eigenschaften und bilden so die Brücke zwischen der Operationalisierung und der anschließenden Bewertung der Prüfungsleistung.

3.6 Gütekriterien und ihre Wechselwirkungen

Die Gütekriterien verdeutlichen, wo die Grenzen eines möglichen Erkenntnisgewinns zu sehen sind. Daher ist es ratsam, jede Prüfung hinsichtlich ihrer Leistungsfähigkeit zu evaluieren. Grotjahn, (2013) fordert, „Sprachtests im Sinne der pädagogisch-psychologischen Diagnostik sollten die folgenden Qualitätsmerkmale aufweisen [...]: Erfüllung der klassischen Gütekriterien der Objektivität, Reliabilität und Validität" (Grotjahn, 2013, p. 211) und verweist darauf, dass diese Kriterien nicht nur bei formellen Prüfungen zu gelten haben, sondern auch bei informellen Tests eine wichtige Rolle spielen.

Da die Ergebnisse von Tests und Prüfungen eine wichtige Rolle bei vielen praktischen Entscheidungen, z.B. Weiterversetzung in der Schule, Beförderung, Einstellung, sprachpolitische Entscheidungen etc., spielen können, formulierte Bachmann schon 1990: „The more important the decision, in terms of its impact upon individuals and programs, the greater assurance we must have that our test scores are reliable and valid“ (p. 78). Im Folgenden werden wichtige Gütekriterien diskutiert:

3.6.1 Validität

Validität (Gültigkeit) ist das wichtigste Gütekriterium einer Prüfung. Man kann damit u.a. einschätzen, ob wirklich das erfasst wurde, was erfasst werden sollte. Der Begriff „Validität“ hat eine wissenschaftshistorische Entwicklung durchlaufen (Kecker, 2011, p. 18 f.) und sich in verschiedene Ausprägungen entwickelt. Für die Praxis der Sprachprüfungen ist die sogenannte „Konstruktvalidität“ das zentrale Konzept von Validität. Das Konstrukt (siehe 3.2) lässt sich nicht direkt beobachten, sondern benötigt Indikatoren, die erfasst werden können. Das Item aus PAT 7.2 unter 3.3 kann bezüglich seiner Validität eingeschätzt werden, denn das Konstrukt, das erfasst werden soll, ist grammatische Kompetenz. Das Item erfasst jedoch die Vertrautheit mit Redewendungen, eventuell auch die Vertrautheit mit dem Testformat, und ist daher nicht valide. Schwieriger ist es, wenn z.B. kognitive Verabeitungsprozesse bei der Bearbeitung von Testaufgaben, das Hörverstehen oder Leseverstehen erfasst werden sollen. Diese sind nicht direkt beobachtbar. Zur Bestimmung der Validität einer Leseverstehensaufgabe muss gefragt werden, ob man auf Grund der beobachteten Leistung in der Prüfung gültige Aussagen zur Leseverstehenskompetenz der Prüflinge in bestimmten realen Situationen formulieren kann.

Eine weitere Form der Validität ist die Augenscheinvalidität, die als Gültigkeit des Tests in den Augen der Getesteten und

Testabnehmer beschrieben werden kann. In der Praxis ist die Validierung einer Prüfung recht anspruchsvoll und kann nur ansatzweise von Lehrenden für informelle Prüfungen durchgeführt werden. Hilfreich ist das Gespräch mit KollegInnen, um möglichst genau zu spezifizieren, was gemessen werden soll, warum eine bestimmte Aufgabe in der Prüfung verwendet wird und warum diese Aufgabe wie bewertet wird. Weiter kann Validität in *interne* und *externe* Validität unterschieden werden. Die interne Validität bezieht sich auf die Eindeutigkeit, mit der die Ergebnisse interpretiert werden können. Sind die Resultate auf die untersuchte Kompetenz zurückzuführen, oder gibt es alternative Erklärungen für das in der Prüfung gezeigte Verhalten? Je mehr Alternativerklärungen möglich sind, desto geringer ist die interne Validität einzustufen. Die externe Validität hingegen bezeichnet die Verallgemeinbarkeit der Ergebnisse über die spezifische Prüfungssituation hinaus. Wenn ein Prüfling eine gute Leistung im Hörverstehen des TestDaF zeigt, wird er diese Leistung dann auch im Studium bei Vorlesungen und Seminaren zeigen können? Für interne und externe Validiät einer Prüfung sind die Konstrukteure der Prüfung zuständig, allerdings sollten Testanwender in der Lage sein, interne und externe Validität kritisch zu reflektieren.

3.6.2 Objektivität

Zentrale Voraussetzung für Objektivität ist die Standardisierung (Vereinheitlichung) der Durchführung und Bewertung einer Prüfung. Geschlossene Aufgabenformate wie *multiple choice* Aufgaben, Zuordnungsaufgaben und Alternativantwort Aufgaben können völlig objektiv bewertet werden. Jedoch besonders bei produktiven und komplexen Kompetenzen (Schreibfertigkeit, sprachliche Studierfähigkeit etc.) ist es schwierig, Objektivität zu gewährleisten. Objektivität wird in die zentralen Konzepte „Durchführungsobjektivität“ und „Bewertungsobjektivität“ unterschieden. Da der vorliegende

Beitrag sich nicht mit der Bewertung beschäftigt, bezieht sich die hier erörterte Objektivität allein auf die Durchführungsobjektivität. Die Durchführungsobjektivität hängt vor allem von zufälligen oder systematischen Abweichungen im Verhalten von Prüfern oder Kommunikationspartnern ab, da diese ihrerseits das sprachliche Verhalten der Prüflinge beeinflussen. Die Hörverstehensaufgabe in DSH Prüfungen kann von unterschiedlichen Prüfern vorgelesen werden, deren Aussprache, Lesegeschwindigkeit, dialektale Färbung, Stimmhöhe, Pausenverhalten etc. jedoch nicht standardisiert ist. Daher hat der Hörverstehensteil in der DSH eine geringere Objektivität als der Hörverstehensteil im TestDaF, denn dort wird an festgelegten weltweiten Prüfungsterminen ein identischer Hörtext digital über Kopfhörer präsentiert.

Es ist jedoch zu beachten, dass Gütekriterien untereinander Wechselwirkungen zeigen. Im TestDaF wird beispielsweise der mündliche Ausdruck in einer so stark standardisierten Weise geprüft (festgelegte Zeiten zur Planung und Produktion der Äußerung, Anweisungen mit Pieptönen, Konversation mit einem Computer oder Tonband etc.), dass fraglich ist, ob hier nicht die Vertrautheit mit dem Prüfungsformat mitgetestet wird, was eine geringere interne Validität bedeutet, und auch ob das in einer so kontrollierten Situation gezeigte Verhalten in einer realen Situation außerhalb der Prüfung gezeigt werden kann; ergo sinkt auch die externe Validität. Am Beispiel des PAT 7.2 kann man sehen, dass eine *multiple choice* Aufgabe zwar hochgradig objektiv ist, damit jedoch nicht automatisch auch eine Verbesserung der Validität erreicht wird. Problematisch hinsichtlich der Durchführungsobjektivität sind des Weiteren Paarprüfungen, die beispielsweise im Zertifikat B1 des Goethe Instituts verwendet werden. Der mündliche Ausdruck hängt stark von der Qualität der Äußerungen des Prüfungspartners ab. Wenn der Prüfling den

Beitrag des Partners nicht verstehen kann, weil dieser nicht auf dem entsprechenden Niveau ist, so kann der Prüfling nicht zeigen, welche mündliche Kompetenz er besitzt. Ferner wird auch das Hörverstehen erfasst, welches doppelt erhoben wird, da die Prüfung das Hörverstehen auch in einem weiteren Prüfungsteil separat erfasst. Die interne Validität und Durchführungs-objektivität sind damit zwar geringer, andererseits ist von einer hohen externen Validität auszugehen, da die Prüfungs-situation einer realen Situation stark ähnelt.

3.6.3 Reliabilität

Das Gütekriterium Reliabilität bezieht sich auf die Zuverlässigkeit, mit der gemessen wird und soll sich kritisch mit der Frage beschäftigen, welcher Anteil der vom Prüfling gezeigten Leistung sich auf Messfehler und wieviel sich auf die intendierte sprachliche Kompetenz zurückführen lässt. Theoretisch sollte beispielsweise ein Lerner, der den strukturell gleichen Test innerhalb kurzer Zeit mehrmals wiederholt, der also theoretisch den gleichen Kenntnisstand besitzt, in den wiederholten Tests das gleiche Ergebnis wie beim ersten Durchlauf erzielen. Ein Kandidat, der den TestDaF nach kurzer Zeit ein zweites Mal ablegt, sollte also auch ein fast identisches Ergebnis erhalten. Gerade bei informellen Tests und Prüfungen wird oft keine hinreichende Reliabilität gewährleistet. Doch auch der sogenannte „DSH Tourismus“, bei dem ausländische Studienbewerber die DSH an verschiedenen Universiäten probieren, weil einige DSHs als „leicht“ gelten, ist ein klares Indiz für mangelhafte Reliabiliät der formellen DSH Prüfung.

Die Reliabilität in der Durch-führungsphase wird beispielsweise durch nicht eindeutige Arbeitsanweisungen beeinträchtigt. Die mehrfach zitierte Aufgabe aus dem PAT 7.2. formuliert „wählen Sie die beste Antwort“, wobei die Formulierung problematisch ist, denn es suggeriert, dass mehrere Antworten richtig sind. Darüber hinaus kann es zu Problemen bei der

Bewertung führen, da Bewerter unterschiedliche Vorstellungen davon haben können, was die beste Antwort ist. Weitere Probleme bei der Reliabilität können durch nicht standardisiertes Material (unterschiedlich lange oder schwere Texte, unterschiedliche Sprecher, unterschiedliches Verhalten von Testern), mangelhafte Messapparatur (mangelhafte Tonbänder, schlechte Druck - oder Kopierqualität, Probleme der Raumakkustik etc.) und eine nicht standardisierte Durchführung (unterschiedliche Dauer zur Bearbeitung von Aufgaben, unterschiedliche mündliche Arbeitsanweisungen etc.) entstehen. Auch hier ist zu bemerken, dass sich ein Maximum an Reliabilität negativ auf andere Gütekriterien auswirken kann. Gerade bei Prüfungen zu mündlichen Kompetenzen ist es schwierig, das Verhalten der Tester oder die Dauer von Redebeiträgen zu standardisieren, da eine solche Beschränkung in aller Regel nicht authentischen Situationen entspricht und damit die Validität beeinträchtigt. Eine maximale Reliabilität ist andererseits kaum praktikabel, wenn es sich um informelle kleinere Tests und Prüfungen handelt

3.6.4. Ergänzende Gütekriterien

Die drei genannten Gütekriterien sind zentral für die Einschätzung der Qualität eines Verfahrens. Insbesondere aus der Prüfungspraxis haben sich eine ganze Reihe weiterer Gütekriterien gebildet, die hilfreich für die Einschätzung eines Verfahrens sind. Der bereits angesprochene Punkt der *Praktikabilität* ist ein solches Gütekriterium, denn wenn eine Prüfung bestimmte Ressourcen erfordert, die nicht vorhanden sind, ist er nicht praktikabel. Man denke hier an online-Prüfungen, die Internetzugang und bestimmte Programme voraussetzen, die nicht überall verfügbar sind. Aus der Psychologie stammt das Gütekriterium der *Ökonomie*, das ähnlich wie Durchführbarkeit danach fragt, welche Voraussetzungen gegeben sein müssen, aber auch konkret den Nutzen im Verhältnis zum entstehenden Aufwand (Mitarbeiter, Raummieten, Gehälter,

Kosten der Apparatur etc.) erfragt.

Das Kriterium *Authentizität* befasst sich mit der Übereinstimmung der Prüfungssituation und Aufgaben mit dem zielsprachlichen realen Verwendungskontext. Insbesondere Prüfungen, die Kompetenzen für spezifische Situationen und Kontexte erheben (bspw. der TestDaF für den studentischen Kontext, Prüfungen zum Wirtschaftsdeutsch, Deutsch für den Tourismus etc.), müssen sich an diesem Kriterium orientieren. Wichtig sind hier vor allem die sprachliche und die situationelle Authentizität. Im Fremdsprachenunterricht spricht man von einer „gemäßigten" Authentizität (Bolton, 1996, p. 21), bei der vor allem die Textmerkmale mit authentischen Texten übereinstimmen, die aber durchaus vereinfacht und bearbeitet sein können, um insbesondere in der Grundstufe eingesetzt werden zu können.

Fairness erfasst die Gerechtigkeit einer Prüfung. Wenn Prüflinge nicht auf Grund höherer Kompetenzen, sondern auf Grund von individuellen Vorteilen während der Prüfung besser abschneiden, handelt es sich um Messfehler. Dies kann z.B. den Abstand zur Hörquelle betreffen, aber auch unterschiedliche Vertrautheit mit dem Testformat. Bei Sprachlehrangeboten mit e-learning oder Fernlernen und obligatorischer Teilnahme am Präsenzunterricht ist auch zu beachten, dass Präsenzunterricht, der in unterschiedlichem Maß zugänglich ist, die Fairness einer Prüfung beeinträchtigt.

Mit dem Gütekriterium *Washback -Effekt* wird erfasst, inwieweit Prüfungen Rückwirkung auf die Unterrichtspraxis (Cheng et al, 2004) oder in einem erweiterten Sinne auch auf Curricularentwicklung oder Zulassungspolitik von Universitäten haben. Für Lehrende ist dies nur im Zusammenhang mit der Unterrichtspraxis nützlich. Wenn beispielsweise Prüfungen nicht das Hörverstehen erfassen, werden die Lernenden den Hörverstehensaufgaben im Unterricht weniger Beachtung schenken und sich stattdessen eher auf die Inhalte konzentrieren, die in Tests und Prüfungen

erfasst werden.

Die *Transparenz* bezieht sich darauf, dass wichtige Merkmale einer Prüfung zu Zielsetzung, Aufgaben, Schwierigkeitsgrad, Strukturellem Aufbau und Ablauf, Bewertungskriterien etc. offengelegt werden sollen. Eine hinreichende Transparenz ist beispielsweise Voraussetzung, wenn die hier vorgestellten Punkte von Testanwendern auf eine Prüfung bezogen werden sollen, um deren Qualität einzuschätzen. Ferner ist Transparenz vorauszusetzen, wenn die Übereinstimmung einer Prüfung mit dem GER oder curricularen Vorgaben eingeschätzt werden soll. Für Prüflinge bedeutet Transparenz auch eine erhöhte Fairness, da insbesondere die Offenlegung der Bewertung einer Prüfungsleistung eine entsprechende Vorbereitung ermöglicht.

Das Gütekriterium der *Trennschärfe* ist vor allem für die Bewertung relevant. Ziel von einzelnen Aufgaben und Items ist es, dass weniger kompetente Prüflinge diese Aufgaben häufiger nicht lösen können als kompetentere Prüflinge. Sind die Aufgaben zu leicht, werden sie von allen Prüflingen gelöst und machen es damit unmöglich, Leistungen in eine Reihenfolge zu bringen. Gleiches gilt für zu schwierige Items. Wenn kein Prüfling diese lösen kann, ist der Erkentnisgewinn minimal, da keine präzisen Aussagen über individuelle Leistungsniveaus innerhalb einer Gruppe möglich sind.

Die *Nützlichkeit* eines Verfahrens kann schließlich als ein übergeordnetes Kriterium verstanden werden, das Kriterien wie Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Wirkung („impact“) und Praktikabilität subsumiert (Bachmann & Palmer, 1996) und deren Gewichtung thematisiert (Grotjahn & Kleppin, 2015).

4. Einschätzung von Sprachprüfungen

Die unter 3. vorgestellten Kriterien verstehen sich als Handreichung zur Einschätzung der Qualität und Eignung von Tests und Prüfungen. Es wurde bereits darauf hingewiesen, dass es unter den Gütekriterien

Wechselwirkungen gibt (siehe 3.6) und es daher nicht Ziel sein kann, alle Gütekritereien maximal zu erfüllen. Vielmehr geht es darum, die Prüfung kritisch reflektieren und entsprechend den jeweiligen Bedürfnissen beurteilen zu können. Dies bedeutet auch, bewusst zu entscheiden, wie wichtig einzelne Gütekriterien in einer spezifischen Prüfung jeweils sind, denn Testanwender haben in der Praxis mit einer Reihe von Beschränkungen zu rechnen, die den Einsatz der theoretisch „besten“ Prüfung verhindern. Technische, organisatorische oder finanzielle Beschränkungen wurden bereits unter „Praktibilität“, „Ökonomie“ und „Nützlichkeit“ angesprochen. Darüber hinaus ist auch zu beachten, dass große Prüfungen mit sehr vielen Prüflingen gewisse Abstriche in der Qualität hinnehmen müssen. Kecker meint, „insbesondere wenn große Kandidatengruppen zur gleichen Zeit geprüft werden sollen, wird beispielsweise in den produktiven Teilkompetenzen die Anwendung direkter Methoden der Kompetenzerfassung durch die Rahmenbedingung erschwert und eine zuverlässige Messung der Fähigkeit sowie deren zeitnahe Auswertung zur Herausforderung“ (Kecker, 2011, p. 21). Dies kann bedeuten, dass beispielsweise produktive Fertigkeiten mit *multiple choice* erhoben werden, z.B. in der Form, die richtige Reaktion auf einen Stimulus zu wählen. Bei einer solchen Prüfung wird die Praktikabilität dann auf Kosten der Validität stärker gewichtet. *Multiple choice* Aufgaben mögen in der Herstellung aufwändig sein und für produktive Kompetenzen nur bedingt valide. Die Beliebtheit dieses Aufgabenformats erklärt sich jedoch aus der sehr einfachen Handhabung für Testanwender. Antworten sind eindeutig richtig oder falsch und die Bewertung kann mit Hilfe eines Lösungsschlüssels in kürzester Zeit vorgenommen werden. Mit Hilfe der hier vorgestellten Überlegungen sollte jedoch ein Testanwender in der Lage sein, Vor- und Nachteile abzuwägen und einen geeigneten Test zu wählen oder selbst zu erstellen.

5. Schlussbetrachtung

Die hier vorgestellten Punkte sollen Lehrende in die Lage versetzten, Prüfungen kritisch zu beurteilen. Dabei liefern die genannten Punkte eine Basis, die jedoch um weitere Punkte erweitert und weiter konkretisiert werden sollte. In der Prüfungspraxis wäre es denkbar, dass die Beurteilungen von Prüfungen im World Wide Web geteilt werden könnten, um auch die Evaluation einer Prüfung für Standardsituationen zu ökonomisieren. Darüber hinaus würde die Anwendung der Punkte 3.1 bis 3.6 helfen, wissenschaftliche Standards für die Prüfungspraxis umfassender zu etablieren und so auch eine Vergleichbarkeit von Sprachprüfungen zu ermitteln. Die mit Hilfe dieses Beitrages mögliche Beurteilung ist keinesfalls vollständig. Die für eine umfassende Evaluation und Validierung von Prüfungen erforderlichen Maßnahmen sind jedoch in der Praxis kaum zu leisten; so erfolgt beispielsweise eine Validierung des TestDaF im Rahmen einer Promotion (Kecker, 2011). Für eine solide Entscheidung über die Eignung einer Prüfung sollte dieser Beitrag jedoch eine brauchbare Handreichung sein.

Eine umfassende Anwendung der Kriterien auf Sprachprüfungen mit zahlreichen Beispielen und auf die Erstellung von Prüfungen in der Praxis würde den Rahmen dieses Artikels sprengen, könnte jedoch in einem weiterführenden Artikel anvisiert werden.

References

Albers, H. G., & Bolton, S. (1995). *Testen und Prüfen in der Grundstufe. Einstufungstests und Sprachstandsprüfungen.* Langenscheidt: Berlin.

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

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

Bolton, S. (1996). *Probleme der Leistungsmessung. Lernfortschrittstests in der Grundstufe*. Berlin: Langenscheidt.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *28* (1), 3-13.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing. Research contexts and methods*. Mahwah, NJ: Erlbaum.

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

Fulcher, G. (2004). Deluded by Artifices? The common European framework and harmonization. *Language Testing Quarterly, 1*(4), 253 – 266.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch. Gemeinsamer Europäischer Referenzrahmen, Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel, Niveau A1-A2, B1-B2, C1-C2*. Berlin: Langenscheidt.

Grotjahn, R. (2013). Sprachtests: Formen und Funktionen. In Hallet, W. & Königs, F.G. *Handbuch Fremdsprachendidaktik.* 3. Auflage. Seelze-Velber: Klett.

Grotjahn, R., & Kleppin, K. (2015). *Prüfen, Testen, Evaluieren*. Klett-Langenscheidt: München.

Hallet, W. (2008). Zwischen Sprachen und Kulturen vermitteln. Interlinguale Kommunikation als Aufgabe. In *Der fremdsprachliche Unterricht Englisch*, *93*, 2-7.

Kecker, G. (2011). *Validierung von Sprachprüfungen. Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen*. Frankfurt am Main: Peter Lang.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Cambridge ESOL: Research Notes*, *37*, 2-5.

Perlemann-Balme, M. (2001). Formen und Funktionen von Leistungsmessung und –kontrolle. In G. Helbig, L.Götze , G. Henrici, & H. J. Krumm. *Deutsch als Fremdsprache. Ein internationales Handbuch. Band II*, 994-1006. Berlin: de Gruyter.

Professional Aptidude Test 7.2. (2011). https://www.opendurian.com/exercises/pat72mar54/1/. Zuletzt gesehen 31.3.2018.

Roche, J. (2013). *Fremdsprachenerwerb Fremdsprachendidaktik*. 3. Auflage. Tübingen: Narr Francke Attempto Verlag.

Schnell, R., Esser, E., & Hill, P. B. (1995). *Methoden der empirischen Sozialforschung.* München: Oldenbourg.

Spolsky, B. (2000). Language testing in the Modern Language Journal. *The Modern Language Journal*, *84*, 536-552.

Stevener, J. (2003). Aufmerksamkeit, Automatisierung und Monitoring: zur Forschungsmethodik. *Fremdsprachen Lehren und Lernen*, *32*, 98-114.

TELC. (2018). *Wer wir sind*. https://www.telc.net/ueber-telc/wer-wir-sind.html. Zuletzt gesehen 28.3.2018.

TestDaF-Institut. (2005). *Musterprüfung 1*. Ismaning: Max Hueber.

Traoré, S. (2016). Translation, intercultural communication and German as a foreign language. Accesses, application possibilities, curricular approach. *Ramkhamhaeng University Journal, Humanities Edition, 35*(1), 27-40.