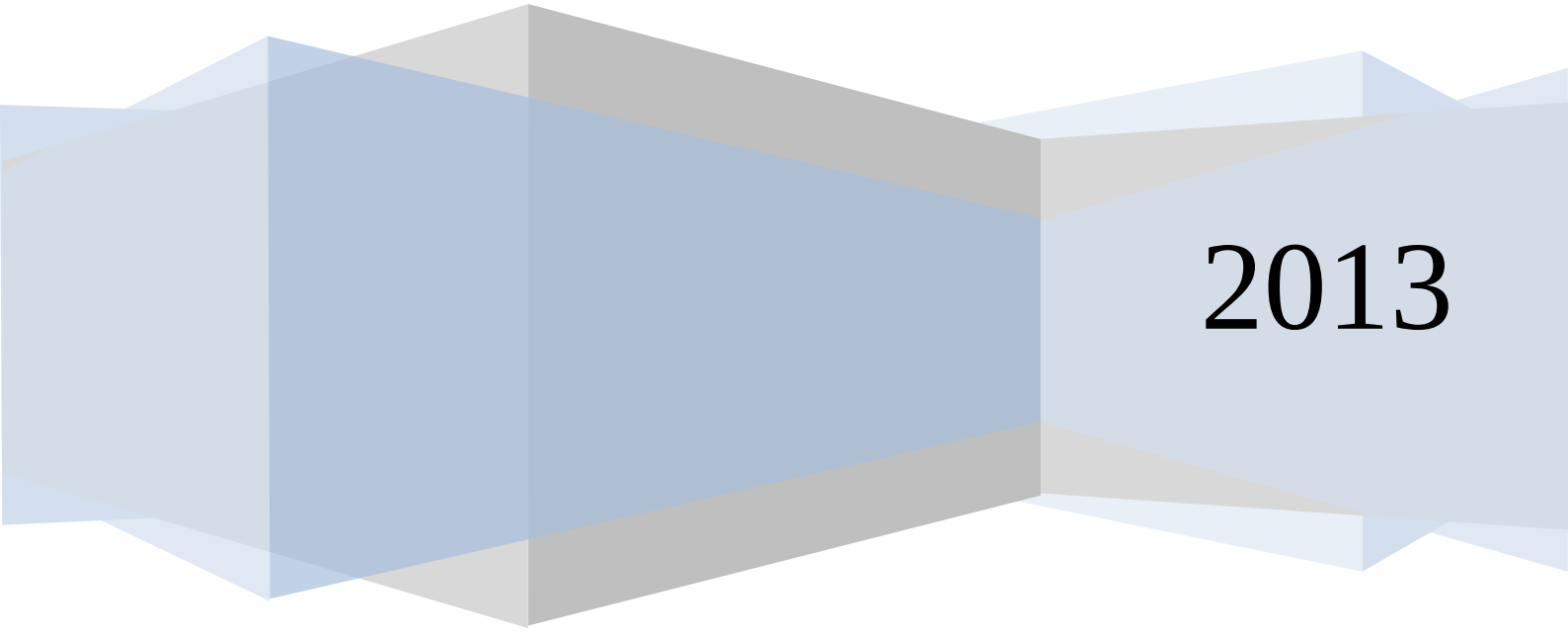
Michael Kranert

Korrigieren, Prüfen und Testen im Fach Deutsch als Fremdsprache

Ein kurzer Leitfaden



2013

Inhaltsverzeichnis

Die Arbeiten an dieser Einführung wurden unterstützt durch ein Rückkehrstipendium des Deutschen Akademischen Austauschdienstes (DAAD), das für dieses Projekt in Zusammenarbeit mit dem Masterstudiengang "Deutsch als Fremdsprache - Kulturvermittlung" an der Freien Universität Berlin von Dezember 2012 bis Mai 2013 gewährt wurde. Mein Dank gilt auch Frau Prof. Hille für die Bereitschaft, mich als Rückkehrstipendiat an der Freien Universität aufzunehmen, und Liane Pohle für das Korrekturlesen und die hilfreichen Kommentare zur ersten Textfassung.

1. Einleitung

Korrigieren und Prüfen sind genuine Aufgaben eines (Fremdsprachen)lehrers, und dennoch scheuen junge Lehrer häufig vor der Rolle des Prüfers zurück, wie Huneke und Steinig (2010) anhand folgender Anekdote illustrieren:

> „Endlich fielen alle Anspannung und die Nervosität der letzten Wochen und Monate von ihm ab, als der frisch gebackene Deutschlehrer nach absolvierter praktischer Ausbildungsphase nun auch das Abschlussexamen bestanden hatte und sich auf den Nachhauseweg machte. Eine aufmerksame Nachbarin empfing ihn daheim mit einer herzlichen Gratulation und einem Päckchen. Er öffnete es: ein Gläschen roter Tinte. Er lächelte höflich und bedankte sich artig, war aber doch etwas konsterniert: In dem Fässchen roter Korrekturtinte wollte er sein Verständnis der neuen Berufsrolle eigentlich nicht symbolisiert sehen.“ (Huneke/Steinig 2010: 223)

Der Sorge des Junglehrers in der Anekdote liegt ein Bild des Korrigierens zugrunde, das den Fehler als Mangel in der Schülerleistung sieht. Hiervon möchte sich natürlich jeder fortschrittliche Pädagoge abgrenzen, schließlich geht es doch um die Förderung der Schüler. Häufig jedoch ist mit der Ablehnung der Rolle des Prüfers auch eine Unsicherheit darüber verbunden, was diese Rolle beinhaltet und wie diese Rolle sinnvoll und gerecht ausgefüllt werden kann.
Dieser Unsicherheit möchte dieser kurze Leitfaden entgegentreten, indem er die Forschungsliteratur zum Testen, Prüfen und Bewerten im Fremdsprachenunterricht knapp und übersichtlich darstellt. Dazu soll zuerst erarbeitet werden, was ein Fehler eines Fremdsprachenlerners ist und welche Rolle die Fehlerkorrektur spielen sollte. Daraufhin werden wir uns theoretisch mit dem Thema „Testen und Prüfen“ auseinandersetzen, indem wir eine Testtypologie, Qualitätskriterien für Leistungstests und Aufgabentypen in Sprachprüfungen diskutieren. In diesem Zusammenhang sollen auch Benotungssysteme und ihre Abhängigkeit von der jeweiligen Kultur zur Sprache kommen – dies ist wichtig, um wenigstens diesen Teil des Kulturschocks für Deutschlehrer im Ausland etwas zu mildern.
In einem zweiten großen Teil wird dargestellt werden, wie die Teilleistungen der sprachlichen Kompetenz – Lesen, Hören, Sprechen und Schreiben getestet werden können.

2. Rote oder grüne Tinte? Fehlerkorrektur im Fremdsprachenunterricht

Unser Lehrer aus der Einleitung lehnt rote Tinte ab. Er will kein kleinkarierter Studienrat sein, der den Schülern ihre Fehler vorhält, sondern sie korrigieren, um ihnen zu helfen. Also stellt sich auch ihm die Frage, was er korrigieren soll. Als Sprachenlehrer muss man sich zuerst bewusst machen, dass alle Sprecher, auch Muttersprachler, von der Norm abweichen, also Fehler machen. So entwickeln beispielsweise Kinder ihre Muttersprache durch das Erproben von sprachlichen Äußerungen. Diese folgen den Regeln, die die Kinder aus Äußerungen ihrer Kommunikationspartner abgeleitet haben. Auch Fremdsprachenlerner zeigen solches

Verhalten. Diese Äußerungen mögen der Norm wiedersprechen, sind aber oft in sich regelhaft, wie eine typische Kinderäußerung zeigt: „Ich habe heute eine Katze geseht." Unsere Sprecherin hier hat möglicherweise die Regelhaftigkeit der Perfektbildung erkannt, aber die unregelmäßigen Formen noch nicht erworben. Dieser **Kompetenzfehler**, der aus einer Regel resultiert, die nicht vollständig erworben wurde, zeigt also gleichzeitig eine Entwicklungsstufe an. Dass es sich um einen solchen Kompetenzfehler handelt, ließe sich an andere Äußerungen der Sprecherin bestätigen, die konsistent dem gleichen Muster folgen.
Anders sieht es bei **Performanzfehlern** aus, die aus einer inkonsequenten Anwendung einer Regel ergeben. Hier würde sich kein klares Bild einer fehlenden Regel oder einer Übergeneralisierung zeigen. Zusätzlich zur Unterscheidung zwischen Performanz- und Kompetenzfehlern sind **Systemfehler** von **Normfehlern** abzugrenzen. Systemfehler sind immer falsch und verstoßen gegen die Normen des lexiko-grammatischen Systems, während **Normfehler** eine stilistische oder situativ unangemessene Äußerung darstellen.
Kleppin (2003) hat in ihrer Einführung in die Fehlerkorrektur folgende Liste typischer Fehlerdefinitionen zusammengestellt, wie sie von Sprachwissenschaftlern und Sprachpädagogen gegeben werden:

> „A Ein Fehler ist eine Abweichung vom Sprachsystem.
>
> B Ein Fehler ist eine Abweichung von der geltenden linguistischen Norm.
>
> C Ein Fehler ist ein Verstoß dagegen, wie man innerhalb einer Sprachgemeinschaft sprich und handelt.
>
> D Ein Fehler ist das, was ein Kommunikationspartner nicht versteht.
>
> E Ein Fehler ist das, was ein Muttersprachler nicht versteht.
>
> F Ein Fehler ist das, was gegen Regeln in Lehrwerken und Grammatiken verstößt.
>
> G Ein Fehler ist das, was ein Lehrer als Fehler bezeichnet.
>
> H Ein Fehler ist das, was ein Muttersprachler in einer bestimmten Situation nicht sagen oder tun würde.
>
> I Ein Fehler ist das, was gegen die Norm im Kopfe des Lehrers verstößt.
>
> J Fehler sind relativ: Was bei einer Lerngruppe in einer bestimmten Unterrichtsphase als Fehler gilt, wird bei einer anderen in einer anderen Phase toleriert." (Kleppin 2003: 19 fff.)

Aus diesen Fehlerdefinitionen hat Kleppin fünf Aspekte einer möglichen Fehlerdefinition destilliert, die ein Fremdsprachenlehrer reflektieren muss (Kleppin 2003: 20ff.):

1 **Korrektheit** (siehe „Systemfehler") - Definition A-C
2 **Verständlichkeit** - Definitionen D und E
3 **Situationsangemessenheit** (siehe „Normfehler") - Definition H
4 **Unterrichtsabhängigkeit** - Definitionen F-I
5 **Flexibilität und Lernerbezogenheit** - Definition J

Das Kriterium der Korrektheit, wie es auch im Begriff des Systemfehlers enthalten ist, ist den meisten Sprachlehrern und -lernern das vertrauteste, schließlich habe wir es alle durch die Kennzeichnung unserer Fehler mit roter Tinte selbst erfahren. Sprachwissenschaftlich aber ist dieses Kriterium hoch problematisch, da unklar ist, auf welchen Standard sich die Korrektur beziehen soll – ist hier immer die aktuelle Ausgabe des letzten Universalwörterbuchs gefragt, die ein aktuelles Bild der deutschen Sprache darstellt, oder sollen regionale Varianten zugelassen werden? Sprachlehrer neigen unbewusst dazu, ihre eigene Sprachvarietät als Maßstab zu nehmen.

Glücklicherweise ist es im Zeitalter des Internets leichter möglich, seine Intuitionen mit Hilfe von Korpora zu überprüfen. Für den Deutschlehrer ist hier das „Digitale Wörterbuch der deutschen Sprache“ (www.dwds.de), entwickelt von der Berlin-Brandenburgischen Akademie der Wissenschaften, ein hilfreiches Werkzeug.

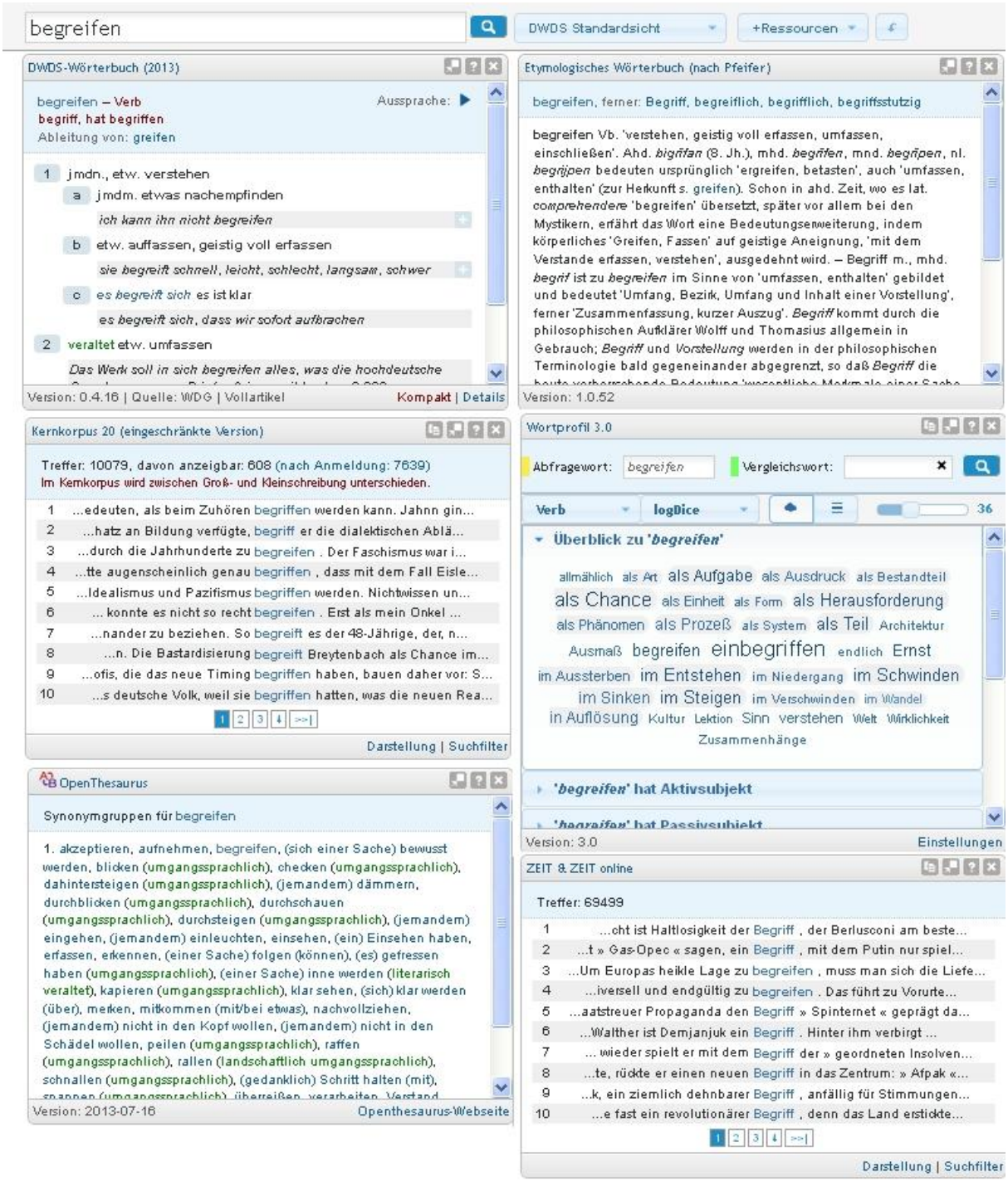


Abbildung 1: Informationen im "Digitalen Wörterbuch der Deutschen Sprache", Screenshot von www.dwds.de

Hier kann man nicht nur auf ein deutsch-deutsches Wörterbuch zugreifen und sich die Etymologie eines Wortes anzeigen lassen, sondern man kann auch kostenlos ein Korpus verwenden, sich also Beispiele aus verschiedenen Kontexten ansehen. Die Website errechnet aus diesem Korpus sogar ein Wortprofil, sodass auch Kollokationen des gesuchten Lexems erkennbar sind. Mit Hilfe dieser deskriptiven Werkzeuge kann der Deutschlehrer seine Intuition überprüfen und verhindern, dass er nur seine Varietät zur Grundlage der Korrektur

macht. Auf der anderen Seite sind solche deskriptiven Normen für den Sprachlerner frühestens ab der Oberstufe handhabbar, ja können in der Grundstufe und Mittelstufe sogar verwirren. Denken Sie nur an die Regel zum Gebrauch von haben und sein im Perfekt/Plusquamperfekt, wo die süddeutsche Variante in der Korrektur/ Prüfung durchaus akzeptabel sein kann, im Unterricht aber verwirrt.
An der Diskussion um Abhängigkeit der grammatischen „Korrektheit" zeigt sich auch, dass Systemfehler und Normfehler nicht scharf zu trennen sind, denn bereits die Varietätendifferenz im Bereich der Perfektbildung ist durchaus auch eine Frage der Situationsangemessenheit. Als Pädagoge ist hieraus zu folgern, dass dem Aspekt der Unterrichtsangemessenheit ein wesentlicher Stellenwert in der Korrektur beizumessen ist und eine differenzierte Fehlerbetrachtung notwendig ist. Um die Schüler beim Erwerb der Perfektbildung zu unterstützen kann es daher in der Grund- und Mittelstufe durchaus angemessen sein, die Süddeutsche Variante in einer Hausaufgabe als falsch zu kennzeichnen, wohlwissend, dass diese Variante existiert.
Ist man sich über das Verständnis der Kategorie „Fehler" und seiner Probleme im Klaren, stellt sich die Frage danach, wie und wozu man als Lehrer korrigiert. Um diese Frage zu beantworten, kann man Korrekturen als Handlungssequenzen analysieren, die sich in Phasen einteilen lassen (Henrici/Herlemann 1986, Kleppin/Königs 1997):

- Eine als korrekturbedürftig angesehene Äußerung wird vom Sprecher selbst oder seinem Gesprächspartner als **Korrekturanlass** wahrgenommen, eine Korrektursequenz wird vom Sprecher (**selbstinitiiert**) oder vom Gesprächspartner (**fremdinitiiert**) begonnen.
- Ein **Korrekturversuch** wird vom Sprecher selbst (**Selbstkorrektur)** oder vom Gesprächspartner (**Fremdkorrektur**) unternommen.
- auf dem Korrekturversuch erfolgt eine **Reaktion,** z.B. eine Bestätigung der Selbstkorrektur durch den Lehrer oder eine Wiederholung der korrigierten Äußerung durch den Lerner.
- in einer **Nachreaktion** kann eine Korrektur zu einem späteren Zeitpunkt wieder thematisiert werden.

Während es sicherlich sinnvoll ist, eine selbstinitiierte Selbstkorrektur im Gespräch zumindest non-verbal zu unterstützen, stellt sich die Frage, ob und wann man eine Fremdkorrektur als Lehrer fremdinitiieren soll. Auch hierauf gibt die Forschungsliteratur Antworten. Von 133 Paderborner Anglistikstudierenden wünschen sich 71,4% eine Korrektur, 51,9% erbaten sich aber eine Korrektur nach Beendigung des Redebeitrags (Gnutzmann/Kiffe 1993: 104). Dass eine Selbstkorrektur einer Fremdkorrektur möglichst vorzuziehen sei, da Selbstkorrekturen dem „natürlichen" Sprachverhalten nahekämen, schlussfolgern Huneke/Steinig (2010: 229f.) aus diesen Zahlen.
Huneke (1995) zeigt außerdem, dass 93,3% der von ihm befragten 196 portugiesischen Germanistikstudenten eine Korrektur ihrer schriftlichen Arbeiten wünschten, allerdings nur knapp die Hälfte dieser Gruppe auch stilistische Korrekturen erwartete.
Dass ein großer Teil der Studenten in der ersten Befragung um eine Korrektur nach ihrem Redebeitrag erbittet, verweist uns auf die methodische Möglichkeit, die fremdinitiierte Fehlerkorrektur ganz in die Nachreaktion zu verlagern: Es ist in der Tat sinnvoll, Korrekturen auf eine bestimmte Unterrichtsphase zu beschränken oder zu konzentrieren, um die Aufmerksamkeit zu fokussieren und auch eine Bewusstmachung von Fehlern zu ermöglichen. Außerdem lassen sich dann auch Lernerstrategien auf einer Metaebene zu thematisieren – z.B. wie man eine selbstinitiierte Selbstkorrektur herbeiführen kann. Zu oft gehen nämlich Fremdsprachenlernende davon aus, dass man Fehler einfach finden kann. In solchen „Fehlersuchphasen" können sie daran herangeführt werden, eigene **Fehlerprofile** zu erstellen

(„Was sind meine drei häufigsten Fehler?“) und mit ihrer Hilfe bewusst nach diesen zu suchen.
Dem Lehrenden gibt ein Fehlerprofil der schriftlichen oder mündlichen Fehler eines Lernenden die Möglichkeit, eine Fehleranalyse durchzuführen, um den Lernenden dann besser beraten zu können. Hierbei lassen sich Fehlerursachen unter anderem nach Performanz- und Kompetenzfehlern unterscheiden und wie folgt klassifizieren:

<table>
<tr><th>Kompetenzfehler</th><th>Performanzfehler</th></tr>
<tr><td>Interferenz (z.B. aus dem Englischen „Ich bin heiß“)</td><td rowspan="4">Unsystematische Fehler, die auf die Beherrschung einer Regel hinweisen, aber aus einer unsystematischen Anwendung resultieren</td></tr>
<tr><td>Übergeneralisierung – Ausweitung einer Kategorie oder Regel auf Elemente, auf die sie nicht zutrifft („Ich gehte in London in die Schule“)</td></tr>
<tr><td>Regularisierung – unregelmäßiges Phänomen wird regelmäßig (Untergruppe der Übergeneralisierung – siehe Beispiel dort)</td></tr>
<tr><td>Simplifizierung – Vereinfachungen wie Gebrauch unflektierter Formen („Ich schreiben Brief“)</td></tr>
</table>

Abbildung 2: Fehlerursachen

2.1.Schriftliche Fehlerkorrektur

Aus den oben zitierten Studien zu Ansichten von Fremdsprachenlernern zu Fehlerkorrektur, die aus verschiedenen nicht statistisch vergleichbaren Lerngruppen stammen und deshalb natürlich nur sehr eingeschränkt verallgemeinerbar sind, lässt sich zumindest die Tendenz ableiten, dass vermutlich der größte Teil aller Fremdsprachenlerner eine Korrektur ihrer schriftlichen Arbeiten wünscht – vermutlich wünschen anteilig mehr Studenten eine schriftliche als eine mündliche Fehlerkorrektur. Pädagogisch scheint eine schriftliche Fehlerkorrektur auch sinnvoller, da hier der Produktionsfluss oder gar ein Dialog nicht unterbrochen wird und diverse Möglichkeiten der Korrektur und Überarbeitung bestehen. Nach der intendierten Nachbearbeitung sollte sich auch die Art der schriftlichen Fehlerkorrektur richten:
Zu Beginn eines Grundstufenkurses oder auch von höheren Kursen, in denen Lernende nach längerer Pause wieder mit dem Fremdsprachenlernen beginnen, ist sicher eine Fehlermarkierung und Berichtigung durch den Lehrer sinnvoll.
Eine Fehlermarkierung durch Korrekturzeichen wird oft in Prüfungen verwendet, kann aber auch, nachdem die Korrekturzeichen in einer Lerngruppe eingeführt wurden, zu einer eingeschränkten Art der Selbstkorrektur führen: die Lernenden kennen die Art des Fehlers und können daher die Regeln noch einmal nachlesen. Folgende Tabelle mit Fehlerkorrekturzeichen (nach Kleppin 2003: 144) kann hierzu verwendet werden, sollten aber auf die Lerngruppe in ihrer Komplexität angepasst werden:

A	Falscher **A**usdruck: Im Gegensatz zur falschen Wortwahl würden hierunter umfassendere Strukturen fallen, wie etwa unidiomatische Wendungen, z. B.: *Wir haben Schwierigkeiten gefunden*, (anstatt: *Wir sind auf Schwierigkeiten gestoßen.)* *Sie machte den ersten Fuß.* (anstatt: *den ersten Schritt)* *Er machte einen Skandal mit seiner Frau.* (anstatt: *Er machte ihr eine Szene.)*
Art	Verwendung des falschen **Art**ikels, z. B.: *Ich mag die Blumen,* (anstatt: *Ich mag Blumen.)* Der erste Satz wäre dann ein Fehler, wenn man sagen möchte, dass man Blumen an sich mag.
Bez	Falscher syntaktischer oder semantischer **Bez**ug, z. B.: *Die Frau arbeitete in der Fabrik seines* (anstatt: *ihres) Mannes. Ich gibt* (anstatt: *gebe) es zu.*
Gen	Verwendung des falschen **Gen**us, z. B.: *Zwischen England und Frankreich liegt nur die* (anstatt: *der) Kanal; der* (anstatt: *das) Kind.*
K	Falscher **K**asus, z. B.: *Ich studiere zwei verschiedenen* (anstatt: *verschiedene) Fremdsprachen. Es gibt einen großen* (anstatt: *ein großes) Problem. Aus religiöse Gründe* (anstatt: *aus religiösen Gründen) ist das nicht möglich.*
Konj	Verwendung der falschen **Konj**unktion, z. B.: *In einem arabischen Land soll ein Mann eine Frau nicht küssen, obwohl sie befreundet* (anstatt: *auch wenn/selbst wenn) sind. Wenn* (anstatt: *als) ich gestern aufwachte.*
M	Falscher **M**odusgebrauch; z. B.: *Wenn ich reich war* (anstatt: *wäre), würde ich nach Deutschland in Urlaub fahren.*
mF	**m**orphologischer **F**ehler, nicht existierende Formen von Verben, Adjektiven und Substantiven, z. B.: *Er grüßt mich mit dröhender Stimme* (anstatt: *dröhnender). Das Gebirge erhebte* (anstatt: *erhob) sich vor mir.*
Mv	Falsches **M**odal**v**erb, z. B.: *Du musst hier nicht rauchen* (anstatt: *darfst).*
Präp	Verwendung der falschen **Präp**osition, z. B.: *Ich kümmere mich über* (anstatt: *um) die Kinder. Er behandelt sie als* (anstatt: *wie) ein Tier.*
Pron	Falscher Pronomengebrauch, z. B.: *Ich frage diesen* (anstatt: *ihn). Ich habe dem* (anstatt: *ihm) geholfen.*
R	Falsche **R**echtschreibung, z. B.: *Sie studirt* (anstatt: *studiert). Wenn Man* (anstatt: *man) jemanden begrüßt, ...*
Sb	**S**atz**b**au: unverständlicher Satz aufgrund mehrerer gleichzeitig auftauchender Fehler, z. B.: *Lehrer fragt Schüler auf Tafel.* (gemeint ist: *Der Lehrer forderte den Schüler auf, an die Tafel zu kommen.)*
St	Satz**st**ellung: falsche Wort- oder Satzgliedstellung, z. B.: *Gestern ich habe* (anstatt: *habe ich) viel gegessen. Ich bin nicht ins Kino gegangen, sondern habe ich* (anstatt: *Ich habe) viel gearbeitet.*
T	Falscher **T**empusgebrauch, z. B.: *Bevor ich esse* (anstatt: *gegessen habe), habe ich mir die Hände gewaschen.*
W	Falsche **W**ortwahl, z. B.: *Er wirft mir einen engstirnigen Blick zu* (anstatt: *skeptischen). Ich wollte Geld gewinnen* (anstatt: *sparen). Das ist gewöhnlich bei uns* (anstatt: *üblich).*
Z	Falsche oder fehlende **Z**eichensetzung, z. B.: *Ich weiß__dass ich nichts weiß,* (anstatt: *Ich weiß, dass ich nichts weiß.)*

Abbildung 3: Korrekturzeichen für Fehler Kleppin (2003: 144)

Eine Fehlermarkierung ohne Kennzeichnung der Art des Fehlers und ohne dessen Korrektur zielt auf eine weitest mögliche Selbstkorrektur. Wichtig ist aber, in alle diese Korrekturen immer auch sogenannte **Positivkorrekturen** einzubeziehen, in denen Lernfortschritte oder besonders gute Formulierungen hervorgehoben werden. Dies fördert einerseits die Motivation, führt aber auch zur Verstärkung von „zufällig“ gefundenen positiven sprachlichen Formulierungen.

Bei der Fehlerkorrektur in Prüfungen, soweit denn tatsächlich eine fehlerorientierte Bewertung vorgenommen wird (siehe Abschnitt 5.2, Seite 40), sollten Fehler gewichtet werden. Es sollte nur als Fehler gewertet werden, was tatsächlich Kursinhalt war und daher vom Lerner auch hätte richtig gemacht werden können. Wie später gezeigt wird, wird von der fehlerorientierten Bewertung heute mehrheitlich abgesehen, da sie nicht die eigentlichen sprachlichen Qualitäten eines Textes erfasst.

2.2.Mündliche Fehlerkorrektur

Im mündlichen Bereich sind die Möglichkeiten der Fehlerkorrektur ähnlich weit gefächert wie im schriftlichen. Wie wir gesehen haben, wird eine lehrerinitiierte Fremdkorrektur während einer Äußerung von Lernern eher abgelehnt. Diese sogenannte **Reparatur** durch den Lehrer, der nach erkannter Äußerungsabsicht dem Lernenden eine Formulierungshilfe gibt, unterbricht die Sprachproduktion und kann das Gespräch im schlimmsten Falle zum Erliegen bringen. Auch wenn es keine Pauschalrezepte zur Frage „Wann und wie korrigieren?“ gibt, gibt es noch Alternativen zur fremdinitiierten Reparatur: einerseits kann über verbale oder besser nonverbal **Missbilligung** die Möglichkeit zur Selbstkorrektur eröffnet werden. Hierzu ist allerdings schon eine gute Sprachkompetenz vorauszusetzen und es ist auch immer die Persönlichkeit des Sprechers mit zu beachten: Manche Lerner mögen diese Art des Hinweises, andere werden verunsichert. Gleiches gilt für den **Hinweis** auf Fehlerort und Fehlerart („Achtung, Wechselpräposition!“).
Um gerade im Anfängerunterricht, in dem Missbilligung und Hinweis nicht immer zum Initiieren einer Selbstkorrektur geeignet sind, Lernende nicht durch laufende Korrekturen zu demotivieren, schlägt Kleppin (2003: 89) vor, **korrekturfreie Übungsphasen** in den Unterricht einzuplanen. Im Fortgeschrittenenunterricht sind dann neben den Aufforderungen zur Selbstkorrektur die bereits erwähnten **nachgeschobene Korrekturphasen** sinnvoll. Hierzu kann man z.B. bei Präsentationen Fehler auf einer Folie sammeln und gemeinsam mit der Lerngruppe reflektieren.

3. Testen und Prüfen

Korrekturen von individuellen Lernerarbeiten und -äußerungen dienen im Unterricht vorrangig der Festigung und Verbesserung von erworbenen Kenntnissen und Fertigkeiten. Eine Testsituation kann die Informationen, die ein Lehrer und Lerner aus den Korrekturen ableiten kann, verlässlicher und vergleichbarer machen. Wir testen also unsere Lerner, um ihnen einerseits eine **Selbstevaluation** zu ermöglichen, andererseits kann ein Test den Lehrer verlässlich und vergleichbar über die Sprachfertigkeiten seiner Lerngruppe informieren und damit als Grundlage für die Unterrichtsreflexion dienen. Auf institutioneller Ebene dienen **Einstufungstests** der Eingruppierung von Lernenden in eine Lerngruppe mit einem Lernniveau, das eine größtmögliche Progression fördert. Auf gesellschaftlicher Ebene schließlich führen bestimmte Arten von Tests zu **Zertifizierung** von sprachlichen Fertigkeiten und bereiten damit eine gesellschaftliche **Selektion** vor: Diese Art von Tests beeinflussen letztlich Lebenschancen von Lernern, indem sie bestimmte weitere Qualifikationen, z.B. ein Hochschulstudium, ermöglichen oder verwehren. Damit steigt die Bedeutung des Tests für den individuellen Lernenden erheblich, weshalb für diese Arten von Tests sehr strikte Qualitätskriterien entwickelt wurden.
Dieses Kapitel wird zuerst in diese **Qualitätskriterien** einführen. Daraufhin wird eine **Typologie von Sprachtests** sowie von abstrakten **Aufgabentypen** entwickelt. Dieses Kapitel schließt mit der Diskussion von **Benotungssystemen** und ihrer kulturellen Abhängigkeit.

3.1.Qualitätskriterien für Sprachtests

Als übliche Qualitätskriterien für einen Sprachtest werden in der testtheoretischen Literatur mehrheitlich Objektivität, Reliabilität und Validität genannt.

(1) Die **Objektivität** eines Tests oder einer Prüfung bestimmt, wie unabhängig das Ergebnis des Tests vom Prüfer ist. Im Idealfall der vollständig objektiven Prüfung käme jeder Prüfer bei der Bewertung eines jeden Tests zum gleichen Ergebnis. Um sich diesem Ideal anzunähern, und für einen menschlichen Prüfer und menschliche Prüflinge ist immer nur eine Annäherung möglich, muss die Durchführung und Auswertung der Prüfung sowie die Interpretation der Ergebnisse standardisiert werden. Wenn die Durchführung eines Tests so reguliert ist, dass es für die verschiedenen Testteilnehmer keine Unterschiede mehr gibt – z.B. in der Akustik bei einer Hörverstehensprüfung – ist ein Test **durchführungsobjektiv.** Die **Auswertungsobjektivität** eines Tests wiederum hängt von den Regeln ab, nach denen die Lösungen der Testteilnehmer bewertet werden: ein Lückentext mit einer geschlossene Liste möglicher korrekter Lösungen wäre also vollständig auswertungsobjektiv (vgl.Grotjahn 2000a: 35). Die **Interpretationsobjektivität** einer Prüfung bestimmt die Unabhängigkeit des Prüfungsergebnisses vom interpretierenden Testbenutzer, also z.B. vom Lehrer, von der Zulassungsstelle oder vom Arbeitgeber. Sie ist unabhängig von der Durchführungs- und Auswertungsobjektivität, da sie vom Noten- oder Zertifizierungssystem (siehe Abschnitt 3.4, Seite 16) abhängt. Ein eindeutiger nummerischer Wert beispielsweise, der den einzelnen Testkandidaten in Relation zu allen anderen Kandidaten setzt, würde eine hohe Interpretationsobjektivität garantieren.

(2) Die **Reliabilität** (=Zuverlässigkeit) eines Tests beschreibt die Genauigkeit, mit der ein Test eine Eigenschaft einer Testperson erfasst. Wenn wir also die Lesekompetenz eines Kandidaten erfassen wollen, bestimmt die Reliabilität, wie genau die Lesekompetenz erfasst wird, das Testergebnis also nicht von anderen Einflüssen wie Prüfer und anderen Kandidateneigenschaften abhängt. Von einigen Autoren wird die Objektivität als Teil der Reliabilität gesehen (z.B. Dlaska/Krekeler 2009: 36). Zur Feststellung der Reliabilität eines

Tests kann man die **Paralleltestreliabilität** eines Testes bestimmen, die das Ergebnis des in Frage stehenden Tests mit dem eines anderen Tests vergleicht. Von diesem Vergleichstest muss angenommen werden, dass er die zu erfassende Eigenschaft zuverlässig misst. Die **Retestreliabilität** beschreibt, wie zuverlässig eine Testwiederholung zu demselben Ergebnis kommt. Diese ist allerdings schwierig zu messen, da es nicht sinnvoll ist, bei den gleichen Testpersonen dieselben Items (Testaufgaben) wieder zu verwenden. In diesem Fall wäre es fraglich, ob dieses Testergebnis nicht durch einen Lerneffekt beeinflusst wird. Zusätzlich kann man bei sprachlichen Fertigkeiten davon ausgehen, dass es sich nicht um statische Fertigkeiten handelt, diese sich also in der Zeit zwischen den Messungen verändert haben können. Von der Forschungsliteratur wird durchgängig angenommen, dass mit steigender Anzahl an Items, die eine bestimmte Fertigkeit testen, die Zuverlässigkeit der Messung steigt. Viele kleinere Tests einer Fertigkeit ergeben also ein besseres Bild über das Leistungsspektrum eines Lernenden.

(3) Wenn wir nach der der **Validität** eines Testes fragen, wollen wir wissen, ob der Test tatsächlich die Fähigkeit misst, die wir testen wollen. Dabei ist zuerst die Frage interessant, ob das Verhalten einer Testteilnehmerin im Test tatsächlich auf eine bestimmte Eigenschaft dieser Person schließen lässt. Dies nennt man **Konstruktvalidität.** Im Bereich der **Inhaltsvalidität** spielt der Aufgabeninhalt eine Rolle: Es wird gefragt, ob der Inhalt typisch für eine Fähigkeit auf einem bestimmten Niveau ist. Es ist kein leichtes Unterfangen, die Inhaltsvalidität zu bestimmen. Die gängige Methode besteht darin, Experten zu befragen und ist daher von subjektiven Urteilen abhängig (siehe hierzu Fulcher 1999). Außerdem ist sie schwer verifizierbar, da der Einfluss des Aufgabenformats nicht sauber vom Inhalt getrennt werden kann (Grotjahn 2000a: 39). Die **kriterienbezogene Validität** wird durch die Übereinstimmung mit externen Kriterien bestimmt, beispielsweise der Übereinstimmung mit anderen Tests. Aus der Perspektive der Testpersonen spielt die **Augenscheinvalidität** eine entscheidende Rolle: Ob die Teilnehmer den Test als sinnvoll wahrnehmen ist unter anderem entscheidend dafür, ob sie ihn ernstnehmen, was wiederum das Testergebnis beeinflussen kann.

(4) Zwischen den hier erläuterten Basiskriterien für einen Test besteht erwartungsgemäß ein Wechselbeziehung (Grotjahn 2000a: 43f.): Objektivität und Reliabilität sind notwendig für einen validen Test, d.h. wenn diese Kriterien nicht erfüllt sind, kann der Test nicht valide sein. Daher kann ein Test auch nicht valider als reliabel sein. Objektivität und Reliabilität sind aber keine hinreichenden Bedingungen für die Validität, denn diese wird auch von anderen Bedingungen beeinflusst.
Für das nicht formale Prüfen im Rahmen von Schule und Universität, das heißt für nicht standardisierte Prüfungen, ist wichtig, dass ein Test mit geringer Validität und hoher Reliabilität gut für die Differenzierung zwischen den Leistungen von Prüflingen geeignet ist. Eine hohe Reliabilität ist leichter herzustellen als eine hohe Validität, da die Validität in der Regel durch komplexe Vortestung und statistische Auswertung von Aufgaben sichergestellt werden muss. Es ist mit solchen Prüfungen allerdings nicht möglich, Vorhersagen über spezifische Fertigkeiten der Prüflinge zu machen.

(5) Bachman und Palmer (1996) schlagen eine Erweiterung der klassischen Kriterien für Tests und Prüfungen im Fremdsprachenbereich vor. Einerseits bewerten sie solche Tests nach ihrer **Authentizität**, d.h. danach, ob sie die Sprachbenutzung außerhalb der Testsituation wiederspiegeln. Hierdurch soll einerseits sichergestellt werden, dass vor allem weniger konstruktvalide Tests in der Unterrichtssituation dennoch eine Aussagekraft hinsichtlich der Fertigkeiten von Lernenden im Alltag haben. Andererseits werden authentischere Prüfungsmaterialien und -aufgaben von Testteilnehmer auch als augenscheinvalider

wahrgenommen. Weiterhin bewerten Bachman und Palmer (1996) die Einbeziehung der Testpersonen in den Testprozess als Kriterium der **Interaktivität** mit ein, indem sie die Frage stellen, inwiefern ein Test das Interesse des Testteilnehmers weckt und an sein Vorwissen anschließt. Relevant sind außerdem die **Auswirkungen** eines Tests auf Lernprozesse und gesellschaftliche Entwicklungen bewerten sie ebenfalls als relevant für die Qualität eines Sprachtests.

(6) Dlaska und Krekeler (2009) weisen darauf hin, dass die genannten Standard-Kriterien für Tests sich in der Regel auf formalisierte und standardisierte Sprachtests beziehen. Diese Tests werden mit hohem personellen und organisatorischen Aufwand für große Teilnehmerzahlen entwickelt und durchlaufen einen komplexes Vortestverfahren, das ihre Qualität auf sehr hohem Niveau sicherstellt. Für die Verwendung in der Schule und in anderen Unterrichtssituationen sind sie daher ungeeignet, da dort solche Ressourcen nicht zur Verfügung stehen. Deshalb haben Dlaska/Krekeler (2009) einen Kriterienkatalog speziell für den Unterrichtsgebrauch entwickelt, indem sie die klassischen Kategorien neuen Kategorien unterordnet und an die Anwendbarkeit im Unterricht angepasst haben. Wir haben diese Kategorien in Abbildung 2 zusammengefasst. Die Autoren ersetzen Validität durch die erweiterte Kategorie **Gerechtigkeit**, da die Validität eines Tests oder einer Prüfung für den Unterricht nur mit nicht zu rechtfertigendem Aufwand zu prüfen ist.

Kriterium	Elemente des Kriteriums	Relevante Frage
Gerechtigkeit *Werden keine Teilnehmer benachteiligt?* *Fühlen sich keine Teilnehmer benachteiligt?*	**Transparenz**	*Haben die Testteilnehmer genügend Informationen über den Test?*
	Bewertung	*Ist die Bewertung schlüssig und nachvollziehbar?*
	Reliabilität	*Sind die Messungen konsistent? Lassen sich die Ergebnisse übertragen?*
	Validität	*Sind die Interpretationen der Testergebnisse gültig?*
Rückmeldung		*Erhalten die Testteilnehmer eine Rückmeldung, die Ihnen für das Lernen hilft?*
Auswirkung		*Hat der Test positive Auswirkungen auf das Lernen und die Motivation zum Lernen?*
Aktivität *Ist die Bearbeitung des Tests eine sinnvolle Lernaktivität?*	**Authentizität**	*Bildet der Test die Kennzeichen der tatsächlichen Sprachverwendungssituation ab?*
	Interaktivität	*Werden Sprachkompetenz, das Vorwissen und die Interessen der Testeilnehmer aktiviert?*
	Unterrichts-prinzipien	*Stimmt die Testaktivität mit den didaktischen und methodischen Prinzipien des Unterrichts überein*

Abbildung 4: Qualitätskriterien für Sprachtests im Unterricht: nach (Dlaska/Krekeler 2009)

Auch die Reliabilität an sich ist für einzelne Lehrkräfte oder durchschnittliche Kollegien nicht empirisch zu testen, denn auch hier benötigt man eine komplexe Infrastruktur. Dlaska und Krekeler (2009) schlagen aber eine Reihe von Fragen und Maßnahmen vor, um die Reliabilität von Prüfungen generell zu erhöhen:

- Tests verlängern,
- mehrere Tests durchführen,
- Items untersuchen: Ist Raten möglich? Ist der Schwierigkeitsgrad der Lerngruppe angemessen?

- Test klar erläutern,
- wann immer möglich, mehrere Prüfer/ Bewerter einsetzen. (Dlaska/Krekeler 2009: 50ff.)

Die Kriterien der **Auswirkung** und der **Aktivität** kennen wir bereits von Bachman und Palmer (1996), sie sollen deshalb hier nicht weiter erläutert werden. Aus didaktischer Perspektive ausgesprochen wichtig ist jedoch das Kriterium der Rückmeldung (Dlaska/Krekeler 2009: 57ff.), schließlich ist die Qualität der Rückmeldung eine wesentliche Voraussetzung für maximalen Lernerfolg, denn die Kombination von großer Unterstützung und hohen Anforderungen ermöglicht effektives Lernen. Außerdem ist dies ein Kriterium, das direkt und problemlos durch den Lehrer erfüllt werden kann. Aus Platzgründen soll die Methodik der Rückmeldung hier nicht ausführlich entwickelt werden, als Anregung kann aber die von Dlaska und Krekeler (2009) erstellte und hier leicht modifiziert wiedergegebene Übersicht über eine Klassifikation der Rückmeldung dienen:

Rückmeldungen				
Zeitpunkt	sofort		später	
Perspektive	Lerner	Lerngruppe	Lehrkraft	automatisch
Gegenstand	Prozess		Produkt	
Typ	beschreibend (Darstellung der individuellen Kompetenz)		bewertend (im Hinblick auf Normen)	
Korrektur	indirekt (reine Fehlerkennzeichnung und -klassifizierung)	direkt (mit Verbesserungsvorschlag)	ohne Hinweise	
Kommunikation	schriftlich		mündlich	
Konsequenzen	gering		gewichtig	
Form	Kommentar		Note o.ä.	

Abbildung 5: Rückmeldungen, nach Dlaska/Krekeler (2009: 61)

(7) Nachdem wir nun, ganz pädagogisch positiv gesinnt, über Qualitätskriterien für Tests und Prüfungen nachgedacht haben, wollen wir uns zum Schluss dieses Abschnitts doch noch der problematischen Seite der typischen Fehler beim Prüfen und Bewerten zuwenden, in der Hoffnung, dass die Kenntnis dieser Fehler auch zu ihrer Vermeidung beiträgt. Einige typische unbewusste Tendenzen menschlicher Wahrnehmung bilden hierbei das größte Problem: Einerseits die Tendenz, **Vorurteile** zu bilden, oder besser Vorklassifikationen von Menschengruppen vorzunehmen, und bei Unsicherheit oft aufgrund dieser Vorurteile anstatt der eigentlichen Faktenlage zu urteilen. Das ist ein ganz normaler und nicht unbedingt moralisch zu verurteilender Mechanismus, der aber natürlich die Aussagekraft von Prüfungsergebnissen beeinflusst. Die beste Möglichkeit, dieses Problem zu vermeiden, besteht einerseits in einer **Blindkorrektur**, also der Bewertung ohne Kenntnis der Person. Da dies aber z.B. bei mündlichen Leistungen kaum möglich ist, sollte idealerweise *nicht* der Lehrer einer Gruppe auch der Prüfer sein. Sinnvoll ist auch, die Zahl der Prüfer zu erhöhen – also auf mindestens zwei oder gar drei, abhängig von der Schwere der Auswirkungen der Prüfung – da hierdurch extreme Fehlurteile vermieden werden. Eine zweite problematische,

aber eben auch nicht unbedingt moralisch zu beanstandende Tendenz – schließlich ist auch sie oft unbewusst – stellt die Tendenz zur **Einseitigkeit** dar. Diese kann als Tendenz zur Milde aus Sorge um den Beurteilten oder als Tendenz zur Härte aus anderen Motiven auftreten. Aus Unsicherheit heraus gibt es aber ebenso die Möglichkeit, dass wir entweder zur Mitte oder zu Extremen neigen. Durch eine Tendenz zur Einseitigkeit wird auch der sogenannte **Halo-Irrtum** oder **Überstrahlungsirrtum** verursacht, bei dem besonders positive oder negative Teilaspekte in Leistung oder Verhalten eines Lerners durch Verallgemeinerung auch auf andere Bereiche übertragen werden. Ähnlich einseitig ist der **Sequenz-Irrtum**, bei dem eine verfälschte Bewertung der Leistung einer Testperson durch den Eindruck der vorangegangenen Leistung eines anderen Schülers entsteht – ein Problem, das nicht systematisch zu bekämpfen ist, aber durch ein Bewusstsein des Prüfers reduziert werden kann.
Unsere Wahrnehmung führt uns aber auch durch soziale Mechanismen gerne in die Irre. Hierdurch kann ein **Kollektivirrtum** entstehen, wenn ein Prüfer sich dem Konformitäts- und Normendruck im Kollegium unterwirft.

3.2. Tests im Fremdsprachenunterricht: eine Typologie

(1) Bereits im vorangegangenen Abschnitt sind wir Begriffen begegnet, die versuchen, Testtypen zu kategorisieren. Die bisher wesentlichste Unterscheidung betrifft die Einteilung in **standardisierte** und **nichtstandardisierte Testverfahren**. Standardisierte Verfahren werden nach wissenschaftlichen Gesichtspunkten entwickelt und in Hinblick auf die oben entwickelten Qualitätskriterien an Versuchsgruppen vorgetestet und statistisch ausgewertet. Dies ist zum Beispiel für alle Aufgaben des TestDaF-Instituts der Fall (siehe http://www.testdaf.de/index.php). Die so standardisierten Tests werden dann in **formellen Testverfahren** durchgeführt; sie finden unter definierten und vergleichbaren Bedingungen statt. So werden u.a. in Hörverstehenstests die Hörtexte mitsamt der Aufgabenstellungen, Wiederholungen und Pausen von einem Tonträger abgespielt, anstatt von der Lehrkraft verlesen zu werden, um eine größtmögliche Vergleichbarkeit der Testbedingungen an verschiedenen Orten zu garantieren. **Informell** sind hingegen alle nicht standardisierten und formalisierten Tests, wie sie in der Regel für den Unterrichtsalltag, aber auch im Rahmen der Schul- und Universitätsbildung z.B. Klausuren eingesetzt werden. Diese Tests oder besser Prüfungen können dann nicht eine ebenso hohe Validität und Reliabilität wie formelle standardisierte Tests erreichen.
Formelle und standardisierte Tests werden meist zur **allgemeinen Leistungsmessung** eingesetzt und sind nicht auf ein bestimmtes Curriculum oder auf einen bestimmten Kurs bezogen**,** wohingegen schulische Lernerfolgskontrollen klar **curriculumsbezogen** sind, da sie den Lernerfolg in Hinblick auf ein bestimmtes Lernprogramm erfassen sollen.

(2) Spezifisch für Sprachtests ist die Unterscheidung in Performanztests und Kompetenztests, welche auf Chomskys Unterscheidung zwischen Sprachkompetenz und Sprachgebrauch (*performance*) zurückgeht (Chomsky 1969, 1965: 4). Chomsky postulierte die Notwendigkeit einer Universalgrammatik für den Spracherwerbsprozess. Die Aufgabe eines Sprachwissenschaftlers ist es Chomsky zu folge, von der Sprachverwendung auf die Sprachkompetenz schließen, wodurch Aussagen über eine Universalgrammatik möglich werden. **Kompetenztests** sollen im Anschluss an diesen Ansatz die Sprachfertigkeiten unabhängig von der Verwendungssituation testen. Dem Test liegt dann ein theoretisches Konstrukt zu Grunde, bei dem Sprache als Code verstanden wird, der zu erlernen ist. Die Beherrschung des Codes kann durch das Abprüfen beherrschter sprachlicher Elemente erfolgen.

Performanztests hingegen sollen die Fähigkeit messen, in einer Situation angemessen in einer Fremdsprache kommunizieren zu können. Hierbei wird angenommen, dass grammatische und lexikalische Kompetenz allein keinen effektiven Sprachgebrauch garantiert und eine Verhaltensvorhersage nur in Verbindung mit Performanzfaktoren möglich ist (Grotjahn 2000a: 57). Beide Testverfahren können als **direkte** oder **indirekte** Testverfahren konstruiert werden. Direkte Tests zielen darauf, genau und isoliert nur die in Frage stehende Fertigkeit zu testen, d.h. der Test benötigt kein komplexes Testkonstrukt, um von den Testergebnissen auf Testkriterien zu schließen. Je indirekter ein Test hingegen ist, desto komplexer sind die ihm zu Grunde liegenden Konstrukte. So ist z.B. ein C-Test[1] ein indirekter Test, da von der Lösung der Lückentextaufgaben auf die vier Grundfertigkeiten Hören, Lesen, Sprechen und Schreiben geschlossen wird. Natürlich ist diese Unterscheidung keine diskrete Unterscheidung sondern ein Kontinuum zwischen den Polen direkt und indirekt, wobei es theoretisch kaum einen Test geben kann, der absolut direkt die sprachlichen Fähigkeiten testen kann, denn wir werten immer die Produkte von Handlungen aus, nicht aber die Kompetenzen selbst. Diese sind abstrakt und nicht direkt beobachtbar.

3.3.Aufgabentypen in Sprachtests

Nicht nur die Tests selber, auch die verwendeten Aufgabentypen lassen sich klassifizieren, Hierzu wollen wir hier zwei Modelle verbinden: Einerseits die Dreiteilung der Aufgaben nach Komponenten, wie sie Doyé (1988: 14ff.) vorschlägt, andererseits aber das Modell der unterschiedlichen Offenheit der Aufgaben, wie es z.B. Grotjahn (2000a: 77ff.) diskutiert. Diese beiden Modelle lassen sich zu einer Kreuzreferenz zusammenfassen, wie sie in Abbildung 6 dargestellt wird.

	S-Komponente (Stimulus)	R-Komponente (Reaktion)
Sprache der Komponente	muttersprachlich fremdsprachlich außersprachlich (z.B. Bild)	muttersprachlich fremdsprachlich außersprachlich (z.B. Bild)
offene Aufgaben		Antwortmöglichkeiten weder Tester noch Testperson exakt vorgegeben
halboffene Aufgaben		Antwortmöglichkeiten dem Tester vorgegeben
geschlossene Aufgaben	Gibt Antwortmöglichkeiten vor	Antwortmöglichkeiten dem Tester und der Testperson vorgegeben

Abbildung 6: Komponenten von Testaufgaben und Aufgabentypen

Die **Stimulus-Komponente/ S-Komponente** eine Prüfungsaufgabe umfasst die Aufgabenstellung und die enthaltenen Materialien wie Lesetexte, Hörtexte oder Grafiken. Die

[1] Ein C-Test ist ein fest definierter Lückentest, bei dem mehrere, thematisch verschiedene Kurztexte präsentiert werden. In diesen ist ab dem zweiten Wort des zweiten Satzes bei jedem zweiten Wort die zweite Hälfte getilgt ist. **Lesetipp**: Grotjahn, Rüdiger (2002): Konstruktion und Einsatz von C-Tests. Ein Leitfaden für die Praxis. In Rüdiger Grotjahn (Hrsg.): Der C-Test. Theoretische Grundlagen und praktische Anwendungen. Bochum: AKS-Verl, S.211–225.
Available online at http://homepage.ruhr-uni-bochum.de/Ruediger.Grotjahn/Grotjahn_KonstruktionC-Test_2002.pdf (27/09/2013)

Interpretationskomponente ist der erwartete mentale Prozess, den die Testperson zur Lösung der Testaufgabe durchführen muss. Da sie hier nicht weiter relevant ist, ist sie in der Kreuzreferenz nicht enthalten. Die **Reaktions-Komponente/R-Komponente** ist die erwartete Reaktion, d.h. das Produkt der Arbeit des Prüflings. Die Kreuzreferenz zeigt nun, welche Varianten von Stimulus- und Reaktions-Komponenten es hinsichtlich der Sprache der Komponenten und der Offenheit der Aufgabenstellungen gibt.

(1) Offene Aufgabentypen wie Interviews, Dialoge oder Aufsätze geben sowohl dem Tester als auch der Textperson keine feste Lösung vor, sondern verlangen die Bewertung der Qualität einer freien Äußerung und nähern sich damit der alltäglichen Sprachverwendung an. Allerdings ergeben sich aus der Vielfalt der möglichen Antworten Probleme der Vergleichbarkeit der Leistungen. Die notwendigen Unschärfe von Bewertungskriterien für solche Antworten (siehe hierzu Abschnitt 5.1) verstärkt diesen Effekt und hat zur Folge, dass solche Aufgaben oft eine geringe Durchführungs- und Auswertungsobjektivität haben.

(2) Halboffene Testaufgabe schränken die Menge der möglichen Antworten erheblich ein, wie z.B. bei Lückentexten. Sie erlauben eine klare Definition akzeptabler Antworten und sind daher auswertungsobjektiver.

(3) Bei **geschlossenen Aufgaben** werden dem Tester und Testteilnehmer mögliche Teile einer Lösung vorgegeben:

- Multiple Choice,
- Alternativformen,
- Zuordnungsformen.

Eine hohe Objektivität und Reliabilität dieses Aufgabenformats wird allerdings erkauft durch die Realitätsferne dieser Aufgaben. Außerdem muss die Validität der Aufgaben in der Regel durch eine komplexe Theorie begründet werden.

(4) Eine große Variationsbreite gibt es bei der Wahl des **semiotischen Mediums der Stimulus- und der Reaktionskomponente**: Sowohl die Aufgabenstellung als auch die geforderte Reaktion des Prüflings können **muttersprachlich, fremdsprachlich** oder auch **außersprachlich** sein. Dem Prüfling kann also als Material und Aufgabenstellung ein Text in seiner Muttersprache, der geprüften Zielsprache oder als Bild vorliegen. Dasselbe gilt für das geforderte Produkt: Der Kandidat kann aufgefordert sein, einen Text (und sei es nur ein Satz) in seiner Muttersprache oder in der Zielsprache zu verfassen. Es ist aber auch möglich, ihn um eine grafische Reaktion zu bitten, beispielsweise kann er aufgefordert werden, eine Wegbeschreibung in eine Karte einzuzeichnen.
Wenn wir uns für eine der Varianten entscheiden, müssen wir Fragen der Validität, aber auch praktische Fragen eine bedenken. Eine muttersprachliche Aufgabenstellung in Hörtests kann garantieren, dass der Test nicht durch mangelndes Lesevermögen in der Fremdsprache beeinflusst wird. Ähnliches gilt bei offenen Aufgaben in Hör- und Leseverstehenstests für die die Reaktionskomponente: Will man wirklich nur das Verständnis prüfen, oder auch die Fähigkeit, dieses Verständnis in ein sprachliches Produkt umzusetzen. Wenn der Prüfling beispielsweise einen Hörtextes schriftlich in der Fremdsprache zusammenfassen will, testet man immer auch die Schreibkompetenz mit. Andererseits sind natürlich muttersprachliche S- und R-Komponenten nur bei Testgruppen möglich, die eine gemeinsame Muttersprache haben. Eine Alternative bietet hier vor allem in der Grundstufe der Einsatz einer graphischen R-Komponente: Dahlhaus (1994) beispielsweise demonstriert dies an einem Test des

Hörverständnisses einer Wegbeschreibung durch das Einzeichnen des Weges in eine Karte (Dahlhaus 1994: 115, 154 fff.).

(5) Tests unterscheiden sich nicht zu Letzt auch darin, ob in ihnen **unverbundenen Aufgaben** präsentiert werden, wie dies meist der Fall bei hoch standardisierten Kompetenztests ist. In ihnen beziehen sich die Aufgaben nicht aufeinander. Performanztests hingegen, die eine größtmögliche Einbettung in einen realitätsnahen Kontext verlangen, sind häufig als **integrative Tests** angelegt, d.h. die Aufgaben des Tests beziehen sich aufeinander.

3.4.Der Maßstab, oder: Woran messen wir (sprachliche) Leistungen?

Was für eine Art der Bewertung am Ende einer Prüfung steht, hängt vom Zweck des Testes und seiner Ausrichtung ab. Bereits zu Beginn dieses Kapitels haben wir die möglichen Zwecke eines Sprachtests diskutiert: sie reichen von einer Selbstevaluation bis zur Vorbereitung eines Selektionsprozesses. Abgestimmt auf diese Funktion werden daher verschiedene Bezugsnormen, also die Maßstäbe für Tests und Prüfungen, gewählt. In der Literatur zur pädagogischen Psychologie werden hierzu drei Bezugsnormen herangezogen (Ingenkamp/Lissmann 2008: 63):

- **Soziale Bezugsnorm** – Vergleich mit anderen Lernenden,
- **Individuelle Bezugsnorm** – Vergleich mit früheren Testergebnissen einer Person,
- **Sachliche Bezugsnorm** – Vergleich mit einem Kriterienkatalog: Kompetenzdefinitionen, Lernziele des Unterrichts und Ähnliches.

Wenn in der Literatur zu Sprachtests von **bezugsgruppenorientierten** (Grotjahn 2000a: 97) oder **normorientierten** (Dlaska/Krekeler 2009: 18) Tests die Rede ist, dann beziehen sich diese auf die soziale Bezugsnorm und vergleichen die Ergebnisse zwischen allen Teilnehmern des Tests. Manchmal werden die Ergebnisse auch auf eine Referenzgruppe bezogen. Bei standardisierten Tests wird hierzu eine Eichstichprobe verwendet. Diese Stichprobe ist eine Gruppe von Personen, deren Kompetenz mit Hilfe anderer Tests oder durch Expertenurteile bereits bestimmt wurde. Sie wird nun verwendet, um den Test zu eichen: Die Vergleichsgruppe führt den Test durch und die Ergebnisse werden statistisch ausgewertet. Vereinfacht gesagt kann das Ergebnis dieser Auswertung sein, dass die Teile der Eichstichprobe, die das Niveau B1 im europäischen Referenzrahmen haben, in unserem Test zwischen 20 und 25 von 50 Punkten erreichen. Also wird angenommen, dass reale Testpersonen, die dieses Ergebnis zeigen, das Niveau B1 erreicht haben. Hierzu muss natürlich vorher die Reliabilität und Validität des Tests überprüft worden sein.

Dlaska und Krekeler (2009) halten normorientierte Tests für besonders aussagekräftig, wenn eine große Menge an Tests ausgewertet wird (Dlaska/Krekeler 2009: 19). Normorientierte Tests differenzieren stark zwischen Teilnehmern und benötigen daher stark differenzierende Items. Ein Item, das von allen Teilnehmern vollständig gelöst wird, ist in einem normorientierten Test nicht sinnvoll.
Kriterienorientierte Tests beziehen sich auf die sachliche Bezugsnorm. In der Regel sind curriculumsorientierte Tests in der Schule oder Universität auf einen Kriterienkatalog als Teil des Curriculums bezogen. Aber auch der Test zur Eignung für ein Studium im Ausland (z.B. Deutschkenntnisse für ein Studium in Deutschland – Deutsche Sprachprüfung für den Hochschulzugang DSH) ist ein kriterienorientierter Test. In einem solchen Test stellt es kein Problem dar, wenn alle Teilnehmer die gleichen Ergebnisse erzielen, denn Ziel ist ja nicht die Differenzierung einer Personengruppe nach Grad der Fähigkeit, sondern die Frage, inwiefern ein definiertes Kriterium durch den einzelnen Testteilnehmer erfüllt wird, oder nicht.

Die individuelle Bezugsnorm spielt in der Selbstevaluation die größte Rolle, kann aber auch ein diagnostisches Werkzeug für einen Lehrer sein, der die Frage beantworten will, wie sich ein Lernender entwickelt. Im schulischen Bereich ist sie für die Binnendifferenzierung in Gruppen wichtig. Sie erlaubt Lernenden, in unterschiedlichem Tempo zu lernen, eigene Ziele zu definieren und zu verfolgen. Außerdem kann die individuelle Bezugsnorm ein wesentlicher Motivationsfaktor beim Lernen sein. Nachdem wir nun Tests auch noch nach ihrer Bezugsnorm differenziert haben, können wir zusammenfassend die Eigenschaften formeller und informeller Sprachtests in einer Übersicht von Dlaska/Krekeler (2009) gegenüberstellen:

	formelle Sprachtests	informelle Sprachtests
Bezug zum Unterricht/ Curriculum	normalerweise nicht	Ja
Bestimmung des Testkonstrukts	theoriegeleitet	Curriculumgeleitet
Normierung und Erprobung	ja	Nein
Bezug	häufig normorientiert	kriteriumsorientiert naheliegend
Anzahl der Prüfer und Teilnehmer	mehrere Prüfer viele Teilnehmer	ein Prüfer wenige Teilnehmer
Erstellung	zentral	Dezentral
Merkmalsdimensionen	häufig Kompetenztests	Performanztest naheliegend
Bezug zum Testkonstrukt	häufig indirekt	direkt naheliegend
Konsequenzen für die Teilnehmer	gewichtig	Gering
Entscheidungsgrundlage	ein Test	mehrere Tests
Ziel	Zertifizierung	Lernen

Abbildung 7: Kennzeichen von Testverfahren nach (Dlaska/Krekeler 2009: 32/footciteff.)

Der genaue Aufbau und die Verwendung des Maßstabs unterscheiden sich bei Sprachtests erheblich in der Prüfung der unterschiedlichen Teilfertigkeiten der Sprachbeherrschung. Traditionell geht man davon aus, dass sich die Sprachkompetenzen in produktive und rezeptive Fähigkeiten unterscheiden lassen, die jeweils die gesprochene Sprache und die Schriftsprache betreffen:

	produktiv	rezeptiv
gesprochene Sprache	Sprechen	Hören
geschriebene Sprache	Schreiben	Lesen

Abbildung 8: Sprachliche Grundfertigkeiten

Diese Grundfertigkeiten enthalten jeweils z.T. spezielle Teilfertigkeiten wie beispielsweise die Orthographie bei Schreiben – einen großen Teil der Teilfertigkeiten wie Grammatik und Lexik haben sie jedoch gemeinsam. Da bei Tests der **rezeptiven Teilfertigkeiten** geschlossene und halb-offene Aufgabenformate dominieren, sind diese Tests leichter zu bewerten, denn die eigentliche Bewertung erfüllt hier am einfachsten das Kriterium der Auswertungsobjektivität: Die Antworten sind vorgegeben, daher kann sich das Urteil der

Bewerter nur in Einzelfällen unterscheiden. Um jedoch einen hoch validen und hoch reliablen Test zu erhalten, müssen die einzelnen Aufgaben sehr gut konstruiert und in einem formalen Verfahren vorgetestet werden. Der Maßstab liegt also hauptsächlich in der Konstruktion, der Auswahl und der Gewichtung der Testaufgaben, die wir für die rezeptiven Fertigkeiten in Kapitel 4 diskutieren werden. Schwierig kann auch die Umrechnung von Rohpunkten in die entsprechenden Notensysteme sein, vor allem bei normorientierten Tests, deren Bewertung statistisch nachbearbeitet wird, um trotzt unterschiedlicher Aufgaben eine gleichbleibende größtmögliche Differenzierung und vergleichbare Bewertung zur erhalten.
Für Tests der produktiven Teilfertigkeiten ist dies anders, da hier bezüglich der Reaktions-Komponente (also des Texts oder der Lösung, die der Prüfling produziert) offene Formate vorherrschen. Dies führt zu Problemen der Bewerterobjektivität. Daher liegt hier der Maßstab in der Definition eines Kriterienkatalogs, der unterschiedliche Ausprägungen haben kann. Die verschiedenen Typen dieser Maßstäbe werden wir in Kapitel 5 kennenlernen. Diese Kriterienkataloge können entweder direkt Notenstufen definieren, oder, vor allem wenn die produktive Kompetenz Teil eines Gesamttests aller Kompetenzen ist, Rohpunkte für bestimmte Qualitäten definieren.

Wenn wir durch die Aufgabendefinition und die Kriterienkataloge die Bewertung in Rohpunkten erreicht haben, stellt sich die Frage, auf welcher **Skala** die Gesamtbewertung eines Tests oder einer Prüfung dargestellt werden soll, welche Notenstufen oder Leistungsstufen der Test also kennt. Um die kulturell sehr unterschiedlichen nationalen oder regionalen Bewertungsskalen verstehen und kritisch betrachten zu können, ist ein Grundverständnis von Skalierungen notwendig. Skalen zur Bewertung lassen sich in Nominal-, Ordinal- und Intervallskalen einteilen.
Nominalskalen erlauben den Vergleich von Dingen hinsichtlich eines Merkmals. Hierdurch werden einfache Gruppen wie „bestanden/ nicht bestanden“ oder „männlich/weiblich“ gebildet.
Ordinal- oder Rangskalen bilden die Rangfolge hinsichtlich der Ausprägung eines Merkmals ab. Ingenkamp und Lissmann (2008) geben hier das Beispiel eines Vokabeltests in einer Englischklasse, in der Sven 28 Vokabeln richtig wiedergegeben hat, Olaf 26, Nico 18, Matthias 12 und Udo 9. Diese Zahlen kann man nun in einer Rangfolge von 1-5 abbilden (Ingenkamp/Lissmann 2008: 47):

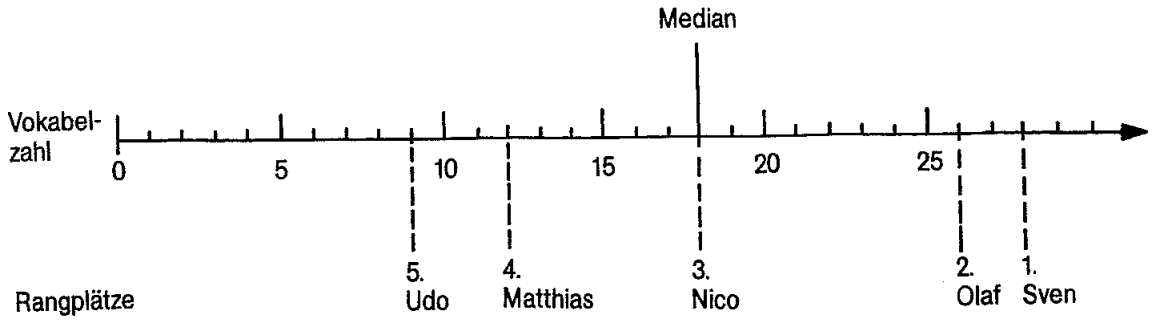


Abbildung 9: Rangfolge der Vokabelkenntnisse einer Lerngruppe; aus:Ingenkamp/Lissmann (2008: 47)

Mit dieser Rangskala kann man etwas über die Reihenfolge der Testergebnisse, hier also die Plätze 1-5 aussagen: Wir wissen, dass Sven auf Rang 1 mehr Vokabeln konnte als Matthias auf Rang 4. Kennen wir, wie z.B. bei den deutschen Notenstufen 1-6 nur diese Rangfolge, nicht aber die Rohpunkte, können wir nichts über die Abstände zwischen den Leistungen sagen, weshalb wir mit den Zahlen einer Rangskala auch keine arithmetischen Operationen wie Addition, Subtraktion, Multiplikation oder Division durchführen dürfen. Daher ist auch der Durchschnittswert solcher Noten, wie er von Lehrern gern errechnet wird, mathematisch

nicht zulässig. Diese Operationen sind nur auf Skalen mit gleichen Abständen möglich. (Ingenkamp/Lissmann 2008: 48). Als zentrale Tendenz auf dieser Art von Skalen gilt der **Median**. Das ist der Wert, über- und unterhalb dessen genau die Hälfte der Stichproben liegt.
Intervallskalen hingegen haben gleich große Abstände: So besteht zwischen 25°C und 50°C sowie zwischen 50°C und 75°C der gleiche Temperaturunterschied. Es ist allerdings nicht möglich zu sagen, dass die Luft bei 50°C doppelt so warm ist wie bei 25°C, da auf der Grad-Celsius-Skala bestimmte Messpunkte – hier 0°C und 100°C – willkürlich definiert wurden, nämlich als Gefrier- und Siedepunkt des Wassers. Außerdem ist auch die Richtung der Skala willkürlich festgelegt. Für die Beschreibung von Proportionen („doppelt so warm“) benötigt man einen natürlichen absoluten Nullpunkt. Das ist beispielsweise beim Gewicht der Fall: „0g“ sagt tatsächlich aus, dass das Merkmal Gewicht hier nicht realisiert ist. Im Bereich der Intelligenzmessung wiederum ist das nicht der Fall, denn man kann eben nicht behaupten, dass jemand, der keine Aufgabe eines Tests lösen kann, keine Intelligenz besitzt. Da eine Intervallskala gleiche Intervalle hat, erlaubt sie arithmetische Operationen und damit auch den Mittelwert.
Verhältnis- und Proportionalskalen schließlich haben einen natürlichen Nullpunkt und gleiche metrische Skalenabstände und erlauben daher alle mathematischen und statistischen Operationen. Solche Skalen gibt es im sozialwissenschaftlichen Bereich nicht, also auch nicht im hierzu gehörenden Bereich der Sprachtests, da es hier keine natürlichen Nullpunkte gibt.

Haben wir nun die Rohpunkte eines Tests, also die Einheiten der Bewertung für bestimmte Aufgaben und Fertigkeiten, sind mehrere Entscheidungen zu treffen. Einerseits muss entschieden werden, in welchem Verhältnis die Rohpunkte in die Gesamtbewertung eingehen. Diese Entscheidung spiegelt in gewissem Maße das Kompetenzverständnis wieder welches der Prüfung zugrunde liegt. Betrachten wir folgendes Verhältnis der Teilprüfungen in einer Bildungseinrichtung für Fremdsprachen:

- Lesen (15%)
- Schreiben (25%)
- Hören (10%)
- Konversation (25%)
- Präsentation (25%)

Hier dominiert das Sprechen mit 50%, der Schwerpunkt der Prüfung liegt also auf der mündlichen Sprachproduktion. Es wäre durchaus möglich, eine Sprachprüfung für akademische Zwecke zu gestalten, deren Schwerpunkt auf der Rezeption schriftlicher Texte liegt.
Weiterhin muss entschieden werden, ob der Test kriterienorientiert oder normenorientiert angelegt ist. Bei kriterienorientierten Tests lässt sich das Verhältnis zwischen Note und Rohpunkten vorher festlegen, da hier nicht die Differenzierung zwischen den Noten das Ziel ist. Bei normorientierten Tests kann dieses Verhältnis entweder erst hinterher festgelegt werden, oder die Trennschärfe der Aufgaben muss vorher an einer Vergleichsgruppe getestet werden, denn bei normorientierten Verfahren möchte man eine möglichst genaue Differenzierung der Leistungen. Daher wäre es problematisch, wenn beispielsweise über die Hälfte der Gruppe die Note 1 erhielten. Handelt es sich bei dem Sprachtest um eine formelle Sprachstandsprüfung, kann die Fragestellung einerseits lauten, ob und wie gut ein Kandidat die Fertigkeiten einer bestimmten Stufe beherrscht, oder der Test kann auf die Differenzierung der Kandidaten nach definierten sprachlichen Fertigkeitsstufen angelegt sein.
Von informellen Sprachprüfungen an Schulen oder Universitäten wird in der Regel erwartet, dass die Note wiederspiegelt, wie gut oder schlecht ein Kandidat einen Kurs abschließt, um damit eine sozioökonomische Selektion wie zum Beispiel die Zulassung zum Studium oder

die Auswahl von Bewerben auf eine Arbeitsstelle, vorzubereiten. Diesen Leistungsfeststellungen liegen **kulturabhängige und historisch gewachsene Bewertungsmuster** zu Grunde. Zum einen sind die Skalen historisch gewachsen und daher häufig schwer vergleichbar: Während die Notenstufen A-F in einigen Ländern dem deutschen Notensystem 1-5/6 ähnlich sind, ist beispielsweise die 100%-Skala in Großbritannien fundamental anders. Das deutsche Bewertungssystem wird sehr häufig als kriterienorientierter „**Mastery Test**" verstanden, in dem ein Teilnehmer mit einem Ergebnis von 100% die Kriterien „in besonderem Maße" erfüllt[2]. Dies wird zwar von dem durchschnittlichen Schüler nicht erwartet, dennoch ist das Erreichen der 100%-Marke möglich. Dass der Leistungstest selbst kriterienorientiert und nicht normorientiert sein muss, ist in der Schulgesetzgebung vieler deutscher Bundesländer festgelegt. So heißt es beispielsweise im Berliner Schulgesetze: „Für die Leistungsbeurteilung maßgebend ist der nach Kriterien des Bildungsgangs festgestellte Entwicklungsstand der Kenntnisse, Kompetenzen, Fähigkeiten und Fertigkeiten der Schülerin oder des Schülers. Die individuelle Lernentwicklung ist zu berücksichtigen." (Senatsverwaltung für Bildung, Wissenschaft und Forschung, Berlin 2004: 58).

Die **britische 100-Prozent-Skala** im Hochschulwesen hingegen funktioniert anders: Hier wird 70% als die Grenze zum „Sehr gut" verstanden, über 80% ist man außergewöhnlich gut, und alles über 90% ist sehr selten. Die 100%-Marke wird also als Grenzwert verstanden, an den man sich annähern, den man aber nie erreichen kann. Hinter diesem System steht die Idee, dass gerade unter den sehr guten Studenten eine starke Differenzierung stattfinden sollte und außergewöhnliche Talente im Notensystem auch als solche ausgewiesen werden müssen. Hinzu kommt, dass das britische System (wie die meisten angelsächsischen Systeme) eine Normenorientierung in der Bewertung bevorzugt. Dies führt dazu, dass Endnoten oft moderiert werden, d.h. die Noten werden am Schluss mathematisch auf einen vorher festgelegten Wert (oft der Durchschnitt mit einer bestimmten Standardabweichung, die für eine weite Streuung sorgt) hin justiert. Das ist natürlich für einen im deutschen System sozialisierten Lehrer ein Kulturschock! Als Begründung für ein solches Vorgehen wird angeführt, dass die Aufgabenschwierigkeit nicht vorgetestet werden kann und variiert, weshalb die Objektivität der Prüfung hinterher hergestellt werden muss. Hierin kann man natürlich einen missverstandenen Anspruch auf Reliabilität sehen, die sich aber eben nicht im Nachhinein herstellen lässt, da sie abhängig von den Testitems und der Bewertungsmethodologie ist. Zusätzlich wird diesen Berechnungen auch noch die Annahme der Normalverteilung von intellektuellen Leistungen zugrunde gelegt, was statistisch bei kleinen Gruppen, natürlich problematisch ist. Außerdem sind diese Gruppen nicht repräsentativ für die Bevölkerung, denn sie sind bereits durch das Zulassungsverfahren zur Universität sowie die individuelle Entscheidung für ein bestimmtes Seminar oder eine bestimmte Vorlesung vorselektiert.

Diese hier dargestellten internationalen und kulturellen Unterschiede in den Bewertungsmaßstäben führen unter anderem zu großen **Problemen bei der internationalen Anerkennung von Leistungen**, da man häufig nur mit Hilfe von Erfahrungen innerhalb des Systems seine Funktionalität genau verstehen kann, und kaum jemand einschlägige Erfahrungen in nur annähernd genügend vielen auch nur europäischen Bildungssystemen hat. Aus diesem Grund sind **Umrechnung von Skalen im Europäischen Rahmen**, wie sie z.B. bei Karran (2005) zu finden sind, äußerst problematisch: So setzt der Autor dort die deutsche Universitätsnote 1,0 mit 96-99 Prozent im britischen System gleich – eine Äquivalenz, die irreführend ist, da die Note 1,0 im deutschen System regelmäßig vergeben wird, die Note 96-99% im britischen System aber so gut wie nie.

[2] Siehe hierzu zum Beispiel für das Land Berlin Senatsverwaltung für Bildung, Wissenschaft und Forschung, Berlin (2004: 58).

Die Vergleichbarkeit der Bewertungssysteme unterscheidet sich auch stark in der Festlegung der Bestehensgrenze, die aber hochrelevant ist, wie Dlaska und Krekeler (2009) zeigen:

> „Bei der Bewertung ist auch die Skala zu berücksichtigen, auf der ein Ergebnis abgebildet wird. Verschiebungen können entstehen, wenn man auf unterschiedlichen Skalen den Schwellenwert unterschiedlich festlegt. Bei einer Prozentskala mit dem Schwellenwert von 50 Prozent differenziert man auch im Nichtbestehensbereich (unter 50 Prozent) noch sehr stark. Das wirkt sich negativ aus, wenn ein Prüfungsteil ausgelassen wurde. Auf einer Notenskala (z.B. 1,0 bis 5,0) liegt der Schwellenwert häufig bei 4,0. Extrem niedrige Ergebnisse in einem Prüfungsteil können leichter ausgeglichen werden." (Dlaska/Krekeler 2009: 47)

Die meisten europäischen Notensysteme sind **unsymmetrisch**, was die Verteilung von Noten oberhalb und unterhalb der Bestehensgrenze betrifft, meistens gibt es mehr Notenstufen und damit Differenzierung oberhalb der Bestehensgrenze (Karran 2005: 9). Außerdem unterscheiden sich die Anzahl und Konditionen **möglicher Wiederholungsprüfungen** in Europa massiv: während in Italien eine Wiederholungsprüfung ohne Nicht-Bestehen der Erstprüfung und ohne Durchschnittsbildung aus beiden Ergebnissen möglich ist (Karran 2005: 10), gibt es in Cambridge keine Wiederholungsprüfungen: ist man erkrankt, kann man entweder einen schlechteren Abschluss hinnehmen, oder das Jahr wiederholen.
Doch nicht nur die Benotungssysteme, auch die Durchführung der Prüfungen und die Erwartungen an Prüfungen variieren erheblich, wie Sullivan (2002) anekdotisch mitteilt:

> „At the anecdotal level, I encountered this issue when confronted with the creation and marking of my .first university examination paper after emigrating to Sweden. As someone educated in Britain I was shocked to discover that the examination was to last six hours. During the examination, I was confronted with differences from the examination setting with which I was familiar. First, students continued to arrive during the first 30 minutes of the examination; second, the background noise level was high with people opening Coke cans, drinking coffee, eating sandwiches and biscuits and, third, smokers are entitled to a smoking break. There was clearly a difference in attitude towards the examination setting. There was also a difference in the students' concept of assessment from the one I was expecting. The students tended to answer the questions with bullet lists rather than as an essay and they were unperturbed that they had finished answering the examination paper in less than the six hours allotted.
>
> The second unexpected aspect of assessment in Sweden was an apparent preference for short-answer questions, which demand the listing of 'facts'. I was shocked when I sat my first business administration paper. The style of question was something that I had not encountered since GCE 'O' level and was an approach to assessment to which I had totally forgotten how to respond in order to achieve a good mark. I failed, in spite of believing that I was well prepared for the examination." (Sullivan 2002: 68f.)

Als mögliche Gründe für das Verhalten der Studenten diskutiert Sullivan (2002), dass in Schweden ein mehr an Zeit zu einer Stressreduktion führe. Außerdem sei die Wiederholungsprüfung ohne Einfluss auf die Note, wodurch den Studenten die Möglichkeit gegeben werde, ihre Leistungen ohne Sanktionen in einem zweiten Versuch zu steigern. Auf empirischem Niveau zeigen Broadfoot et al. (2000) in einer breiter angelegten Untersuchung, dass bereits Grundschüler (hier vergleichend in Großbritannien und Frankreich) sich in ihren Herangehensweisen an Prüfungsaufgaben unterscheiden, denn die lokalen Schul- und Lernkulturen stellen bereits verschiedene Anforderungen an den Lösungsweg schulischer Aufgaben und die Präsentation der Lösungen. Für international tätige Fremdsprachenlehrer bedeutet das gerade in heterogenen Gruppen, dass das **Üben von Prüfungsformaten** für ein faires Prüfungsverfahren notwendig ist und dass man sich, wenn man in anderen Kulturen als

Prüfer tätig ist, man sich sehr genau mit dem neuen System im kritischen(!) Vergleich mit dem eigenen Hintergrund vertraut machen muss.

3.5. Der Gemeinsame europäische Referenzrahmen (GeR) als Maßstab für Sprachtests

Eine Möglichkeit, sich für sprachliche Prüfungen hinsichtlich der Maßstäbe zu orientieren, bietet der „Gemeinsame europäischer Referenzrahmen für Sprachen“ (GeR), welcher im Jahr 2000 vom Europarat vorgelegt und 2001 vom Goethe-Institut gemeinsam mit der Kultusministerkonferenz und dem Österreichischen Sprachdiplom auf Deutsch veröffentlicht wurde (Trim et al. 2009). Der Europarat als eine der ältesten gemeinsamen politischen Organisationen in Europa ist nicht zu verwechseln mit der Europäischen Union. Er wurde 1949 mit dem Ziel gegründet, durch eine engere Zusammenarbeit den sozialen und wirtschaftlichen Fortschritt in seinen 47 Mitgliedsstaaten zu fördern. Seine Abteilung für Sprachenpolitik im französischen Strasbourg engagiert sich seit den 1970er Jahren verstärkt in der Förderung der Mehrsprachigkeit.
Obwohl die mittlerweile weithin bekannten Referenzniveaus A1-C2 als Kern des GeR wahrgenommen wurden, ist der Referenzrahmen mehr als das: er ist ein sprachpolitisches Dokument des Europarats, ein linguistisches Dokument zur Diskussion von Sprachkompetenz und ein sprachdidaktisches Kompedium. Erst in seiner letzten Fassung wurden die ausführlichen Systeme der KANN-Deskriptoren hinzugefügt. (Quetz 2003: 42) Diese Deskriptoren gehen zurück auf das System der ALTE-Levels[3], das bereits ausführlich getestet wurde, wurden aber durch das sogenannte Schweizer Projekt (vgl. North/Schneider 1998) erweitert. Die KANN-Beschreibungen sind wie folgt strukturiert:

> KANN + <was> + <Merkmale des Textes/ der Situation> + VERB + <Bedingungen und Einschränkungen>“ (Quetz 2010: 196)

Hier ein Beispiel aus dem Set „Hörverstehen allgemein“ auf der Stufe A2:

> „Kann Wendungen und Wörter verstehen, wenn es um Dinge von ganz unmittelbarer Bedeutung geht (z. B. ganz grundlegende Informationen zu Person, Familie, Einkaufen, Arbeit, nähere Umgebung), sofern deutlich und langsam gesprochen wird.“ (Trim et al. 2009: 41)

Die Stufen A-C repräsentieren die alten Stufen „Grundstufe“, „Mittelstufe“ und „Oberstufe“ und differenzieren diese weiter:

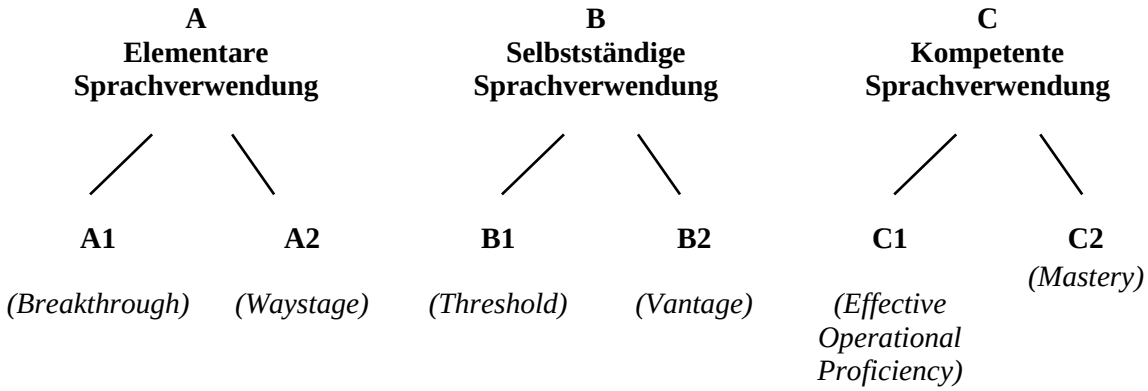


Abbildung 10: Stufensystem des GeR (Trim et al. 2009: 34)

[3] ALTE=The Association of Language Testers in Europe; http://www.alte.org/

Zu allen Stufen gibt es eine Globalskala, die dann in ca. 50 untergeordnete Skalen zu verschiedenen sprachlichen Tätigkeiten, Genres und Mitteln der Sprachverwendung differenziert wird. Am Ende der Kapiteln 4.1 - 5.2 finden Sie jeweils die Grobdeskriptoren für die behandelte Teilkompetenz.
Der GeR stellt damit ein breit ausgefächertes System der Kompetenzbeschreibung für sprachliche Fertigkeiten zur Verfügung, das bei der Entwicklung von Lehrmaterial und Prüfungen als orientierender Standard gelten kann. Er ist allerdings, wie die Autoren selbst betonen, kein festes Curriculum und kein Instrument zur Prüfungsbeurteilung.
Mittlerweile sind viele standardisierte Tests für Fremdsprachen in Europa am GeR ausgerichtet, Abbildung 11 zeigt eine Übersicht über die wichtigsten formellen Sprachstandstests für Deutsch und ihre Orientierung am GeR:

Test	GER	Zielgruppe	Anbieter
Start Deutsch 1, SD1	A1	Erwachsene	Goethe-Institut
Fit in Deutsch 1, Fit 1	A1	Jugendliche	Goethe-Institut
Kompetenz in Deutsch 1, KID 1	A1	Kinder und Jugendliche	Verein Österreichisches Sprachdiplom Deutsch
Grundstufe Deutsch 1, GDI	A1	Jugendliche, Erwachsene	Verein Österreichisches Sprachdiplom Deutsch
Start Deutsch 2, SD2	A2	Erwachsene	Goethe-Institut
Grundstufe Deutsch 2, GD2	A2	Jugendliche, Erwachsene	Verein Österreichisches Sprachdiplom Deutsch
Fit in Deutsch 2, Fit 2	A2	Jugendliche	Goethe-Institut
Kompetenz in Deutsch 2, KID 2	A2	Kinder und Jugendliche	Verein Österreichisches Sprachdiplom Deutsch
Zertifikat Deutsch, ZD	B1	Erwachsene	Goethe-Institut. Verein Österreichisches Sprachdiplom Deutsch, Volkshochschulen
Zertifikat Deutsch für Jugendliche, ZDJ	B1	Jugendliche	Goethe-Institut, Verein Österreichisches Sprachdiplom Deutsch
Deutsch-Test für Zuwanderer, DTZ	B1	Zuwanderer	Bundesamt für Migration und Flüchtlinge
Deutsches Sprachdiplom der Kulturministerkonferenz, DSD I	A2/ B1	Schüler, Aufnahme in Studienkollegs	Auslandsschulen
Zertifikat Deutsch für den Beruf, ZDfB	B2	Beruf	Goethe-Institut, Verein Österreichisches Sprachdiplom Deutsch
Goethe-Zertifikat B2	B2	Erwachsene	Goethe-Institut
Mittelstufe Deutsch, MD	B2	Erwachsene	Verein Österreichisches Sprachdiplom Deutsch
Goethe-Zertifikat C1	C1	Erwachsene	Goethe-Institut
Oberstufe Deutsch, OD	C1	Erwachsene	Verein Österreichisches Sprachdiplom Deutsch
Prüfung Wirtschaftsdeutsch International, PWD	C1	Beruf/Wirtschaft	Goethe-Institut
Deutsches Sprachdiplom der Kultusministerkonferenz, DSD II	B2/ C1	Schüler, Studium	Auslandsschulen
Zentrale Oberstufenprüfung, ZOP	C2	Erwachsene, Studium	Goethe-Institut
Wirtschaftssprache Deutsch, WD	C2	Beruf/Wirtschaft	Verein Österreichisches Sprachdiplom Deutsch
Kleines Deutsches Sprachdiplom, KDS	C2	Erwachsene, Studium	Goethe-Institut

TestDaF	B2 bis C1	Studium	Prüfungszentren weltweit
onDaF	A2bis C1	Einstufungstest	DAAD-Lektorate und Prüfungszentren weltweit
Deutsche Sprachprüfung für den Hochschulzugang, DSH Stufen 1 bis 3	B2bis C2	Studium	Deutsche Hochschulen
Großes Deutsches Sprachdiplom. GDS	über C2	Studium, Berufs-tätigkeit	Goethe-Institut

Abbildung 11: Formelle Sprachprüfungen Prüfungen im GeR, nach Huneke/Steinig (2010: 237f.)

Zusätzlich gibt es seit 2009 auch ein Handbuch des Europarats, wie Prüfungen valide und reliabel auf den GeR bezogen werden können (Council of Europe 2009). Das dort vorgeschlagene Vorgehen ist jedoch ausgesprochen komplex und verlangt neben einem detaillierten Studium aller Deskriptoren eine Testspezifikation in Hinblick auf den GeR, einen Vergleich der eigenen Aufgaben mit standardisierten Beispielen, ein intensives Training der Prüfer an standardisierten Beispielen sowie eine Vortestung und statistische Bewertung der Aufgabenstellungen. Dies ist selbst in sehr großen Teams nur mit großem Aufwand zu leisten und nur sinnvoll für große Gruppe von Studenten. Hinzukommt, dass ein GeR-relevanter Test häufig auch nicht sinnvoll vereinbar mit institutionsinternen Notwendigkeiten der differenzierten Bewertung für Studienleistungen ist, denn eine Kompetenzstufeneinordnung, wie sie der GeR leistet, ist kein Lernfortschrittstest, wie er von Universitäten oder Schulen aufgrund der Selektionsvorbereitung verlangt wird. Außerdem sind in universitären Spracheinrichtungen Lerngruppen häufig aus finanziellen und organisatorischen Notwendigkeiten heraus heterogen, d.h. es gibt nicht ausreichend differenzierte Niveaustufen, weil es für die dafür nötige Anzahl an Lerngruppen nicht ausreichend Studenten gibt. Daher ist also auch ein Niveautest mit festen Eingangsvoraussetzungen in den Kurs problematisch.
Es gibt aber auch wesentlich fundamentaler **Kritikpunkte am GeR**. Einerseits stehen die Verben in den Deskriptoren, die zur Beschreibung der Fertigkeiten verwendet werden, theoretisch nicht auf psycholinguistischer Grundlage und sind deshalb unbrauchbar in der Testerstellung. So fehlt beispielsweise eine Theorie des Textverstehens, wodurch die Validität einer Textverstehensprüfung gefährdet wäre (Quetz 2010: 198, siehe auch Weir 2005). Außerdem wird eingewandt, dass eine Entscheidung des Niveaus von sprachlichem Material oder gar Testaufgaben anhand des GeR kaum möglich sei, da hierzu in der Regel standardisierte Aufgaben als Vergleich herangezogen werden (Quetz 2010: 200). Im GeR gibt es zudem weder Angaben zu möglichen Textsorten in der Rezeption und Produktion auf den verschiedenen Niveaus, noch wird die Schwierigkeit syntaktischer Strukturen reflektiert (Weir 2005: 291ff.). In einer besonders interessanten Studie zeigt Wisniewski (2010), dass bei der Verwendung der GeR-Skala zur Bewertung von Tests und Prüfungen häufig dennoch andere Kriterien bei den Bewertern eine Rolle spielen, wie z.B. die Fehlerzahl. Bewerter lösen Kriterien auch aus dem Kontext der Skalen im GeR heraus und wenden sie auf andere Niveaus an. Dies liege unter anderem daran, dass die Skalen stark fragmentiert und die Deskriptoren vage formuliert seien. Prüfer und Bewerter kompensieren dies mit Alternativstrategien wie fehlerorientierter Bewertung (Wisniewski 2010).

4. Rezeptive Fertigkeiten testen: Lesen und Hören

(1) Die spezifische Schwierigkeit des Testens von rezeptiven Fertigkeiten besteht darin, dass es um den Prozess des Verstehens geht, aus dem Verständnis resultiert. Beides ist nicht direkt beobachtbar, sondern nur die aus dem Verständnis folgenden Handlungen (Grotjahn 2000c: 11). Das Lese- und Hörverstehen sind sich als rezeptive Fertigkeiten grundsätzlich ähnlich, so dass wir die Gemeinsamkeiten hier auch gemeinsam behandeln werden, während wir die Spezifika in den Unterkapiteln besprechen werden.
Zuerst einmal ist festzuhalten, dass man beim Lesen und Hören **Lesestrategien** und **Hörstile** unterscheiden kann (Bolton 1996: 22):

- Detailverstehen: Erfassung jedes Details eines Textes,
- selektives Lesen/Hören: Erfassung bestimmter situationsrelevanter Aussagen,
- globales Lesen/ Hören: Verständnis der Gesamtaussage, Text wird nicht Wort für Wort gelesen.

Für die **Aufgabenschwierigkeit** für Tests rezeptiver Fertigkeiten ist daher nicht der Text entscheidend, sondern die Anforderungen der Aufgaben an Lesestrategien und Hörstile. (Bolton 1996: 24, Pollitt 1986, Buck et al. 1997) Aus Auswahlkriterien lassen sich aber folgende Fragen und Aussagen formulieren (vgl. Bolton 1996: 25, Grotjahn 2000c: 36ff.)

- **Wortschatz**: Wie hoch ist der Anteil an ungewöhnlichen und unbekannten Wörtern? – Die Schlüsselwörter sollten bekannt sein.
- Ist der **Satzbau** stufenangemessen?
- **Inhalt**: Setzt der Text außertextuelle landeskundliche Kenntnisse voraus?
- **Inhaltliche Polyvalenz**: Texte, die verschiedene konkurrierende Lesarten erlauben, sind für Verstehenstests problematisch, vor allem für geschlossene Aufgabenformate, die eine eindeutige Lösung voraussetzen.
- **Authentizität der Aufgaben:** Um eine Generalisierung über die Fähigkeit zur Lösung bestimmter zielsprachlicher Probleme in Performanztests zu erlauben, sind authentische Aufgaben auch in Hinsicht auf die Lesestrategien und Hörstile notwendig. Dies ist beispielsweise relevant für die Position der Fragen zum Text: Authentizität erfordert oft Fragen vor dem Text, da das Leseziel die Informationsentnahme ist.

Da (nicht) vorhandenes **Hintergrundwissen** das Lesenverstehen beeinflussen kann, ist es für das Konzipieren von Prüfungen notwendig, entweder Hypothesen zum gemeinsamen Hintergrundwissen der Kandidaten formulieren oder interkulturelle Kompetenz als Teil des Testkonstrukts mitzubewerten.

(2) Wir wollen nun Charakteristika und Probleme der verschiedenen Aufgabentypen, die in Verstehenstests Anwendung finden, diskutieren.
Offene Aufgaben sind oft Fragen zum Text, die frei und schriftlich beantwortet werden müssen. Sie eigenen sich je nach Fragetyp für das Prüfen des Detailverstehens (wie fast alle W-Fragen) und des Globalverstehen (wie Zusammenfassungen). Wenn die Reaktions-Komponente (also die Antwort) in der Zielsprache verlangt wird, ist allerdings zu beachten, dass die schriftliche Ausdrucksfähigkeit mitgetestet wird. Daher kann das Ergebnis nicht als genaue Vorhersage der rezeptiven Fertigkeiten eines Lernenden interpretiert werden.
Die typische Verstehensaufgabe ist daher häufig eine **geschlossene Aufgabe.** Der Typ der **Multiple-Choice-Aufgabe** dominiert hierbei. In diesen Aufgaben, die eher auf das Detailverstehen gerichtet sind, werden meist drei Antworten vorgegeben, wobei eine Antwort

die richtige Lösung darstellt, während die anderen – falschen – Antworten als **Distraktoren** dienen. Die Aufgaben werden in der Regel so konstruiert, dass auf fünf Zeilen Text etwa eine Aufgabe kommt und die Aufgaben dem Textverlauf folgen, um die Orientierung zu erleichtern. Die Schwierigkeit besteht darin, einerseits eine eindeutig richtige Lösung und mehrere eindeutig falsche Distraktoren zu formulieren, die aber andererseits vom Text her plausibel und möglichst gleichlang sein müssen. Bei der Formulierung sollten Verneinungen vermieden werden und die einzelnen Aufgaben nicht aufeinander bezogen sein. Die Vorteile dieses Aufgabentyps bestehen in seiner Bewertungsökonomie, denn hier sind automatisierte Bewertungen möglich. Auch prüft dieser Aufgabentyp das Verstehen unabhängig von der Schreibfertigkeit. Allerdings ist die Qualität der Wahlmöglichkeiten schwer abzuschätzen, schließlich stellen sich etwa die Hälfte der konstruierten Items in empirischen Tests als unbrauchbar heraus (Grotjahn 2000c: 60). Dieses Problem macht eine empirische Vortestung dieser Aufgaben hinsichtlich ihrer Reliabilität und Validität zwingend notwendig. Das ist allerdings für den Alltagsgebrauch im Unterricht und in schulischen Bildungseinrichtungen unrealistisch, unökonomisch und daher nicht empfehlenswert (Dlaska/Krekeler 2009: 80). Kritisiert wird auch die mangelnde Authentizität dieses Aufgabentyps.

Texte	Aufgabentyp	Art des Leseverstehens
Meinungsäußerungen mehrerer Personen (z.B. in Umfragen, Interviews)	Zuordnungsaufgaben	Global und Detailverstehen
längere geschlossene Texte (z.B. Erzählung, Bericht, Sachtext)	offene Aufgaben	Global-, Detail- und selektives Verstehen
	Multiple-choice-Aufgaben	Detailverstehen
	Alternativ-Antwort-Aufgaben	Global- und Detailverstehen
Texte, die man in realen Alltagssituationen überfliegt, um schnell eine bestimmte Information zu finden (z.B. Fahrpläne, Fernsehprogramme, Veranstaltungskalender)	Zuordnungsaufgaben	selektives Verstehen
	offene Aufgaben	

Abbildung 12: Texte - Aufgaben - Art des Leseverstehens nach Bolton (1996: 38), Übersicht von MK

Bei Fragen mit **Alternativantworten** werden Aussagen vorgegeben, die als richtig oder falsch zu bewerten sind. Hiermit kann man das Global- und Detailverstehen testen. Die hohe Ratewahrscheinlichkeit von 50% kann durch eine Option „nicht im Text“ verringert werden, hier muss aber eine eindeutige Entscheidung zwischen „falsch“ und „nicht im Text“ möglich sein, was manchmal problematisch sein kann. Um dem Problem der Polyvalenz vorzubeugen, ist eine Variante möglich, die eine freie Begründung der Antwort erfordert und daher einem halboffenen Aufgabentyp entspricht. Generell werden Alternativantworten kritisch gesehen, denn es gibt empirische Belege für ihre geringe Reliabilität, Validität und Trennschärfe (Grotjahn 2000c: 62).
Ebenfalls als geschlossener Aufgabentyp gilt die **Zuordnungsaufgabe,** bei der passende Teile, z.B. von Grafiken, Karten, Bildern oder Aussagen bestimmten Textabschnitten zugeordnet werden müssen.

Als **halboffene Aufgaben**, die den starken Einfluss der schriftlichen Ausdrucksfähigkeit sowie das Problem der unklarer Distraktoren vermeiden, gelten **Lückentexte** oder Aufgaben, bei denen Sätze vervollständigt werden müssen. Hiermit wird das Detailverstehen getestet. Möglich sind auch Aufgaben zum **Informationstransfer**, die verlangen, dass Informationen aus dem Text in eine andere Form übertragen werden – z.B. soll ein beschriebener Weg in eine Karte eingetragen werden. Hierbei müssen allerdings die Testinstruktionen besonders sorgsam formuliert werden, damit sie auf den verschiedenen Kompetenzstufen verständlich ist.

Hinsichtlich der **Schwierigkeit der Aufgaben** gilt, dass offene Aufgaben tendenziell leichter sind als geschlossene. Explizite Aufgabenstellungen sind grundsätzlich leichter zu verstehen als implizite (Solmecke 2000: 73). Für ein faires Verfahren ist darauf zu achten, dass alle Aufgabentypen geübt werden müssen, denn dieses Handlungswissen ist eine Voraussetzung für das erfolgreiche Bearbeiten einer Aufgabe (Solmecke 2000: 74). Ob in den Tests ein Wörterbuch zugelassen ist, hängt vom Zweck des Tests und bei curriculumsbezogenen Tests vom Ziel des Kurses ab. Wenn der **Wörterbuchgebrauch** Lernziel ist, dann kann und sollte er auch getestet werden. Außerdem geben Wörterbücher den Lernenden ein Stück Sicherheit in der Prüfung (Dlaska/Krekeler 2009: 91). Eine Veränderung des Textkonstrukts hinsichtlich des Einflusses des Wortschatzes ist allerdings nicht nachweisbar (Alderson 2000: 99 ff.).

Abbildung 13: GeR Deskriptoren für eine Selbstevaluation der Lese- und Hörverstehenskompetenz (Trim et al. 2009: 36)

Verstehen		
	Hören	**Lesen**
C2	Ich habe keinerlei Schwierigkeit, gesprochene Sprache zu verstehen, gleichgültig ob "live" oder in den Medien, und zwar auch, wenn schnell gesprochen wird. Ich brauche nur etwas Zeit, mich an einen besonderen Akzent zu gewöhnen.	Ich kann praktisch jede Art von geschriebenen Texten mühelos lesen, auch wenn sie abstrakt oder inhaltlich und sprachlich komplex sind, z. B. Handbücher, Fachartikel und literarische Werke.
C1	Ich kann längeren Redebeiträgen folgen, auch wenn diese nicht klar strukturiert sind und wenn Zusammenhänge nicht explizit ausgedrückt sind. Ich kann ohne allzu große Mühe Fernsehsendungen und Spielfilme verstehen.	Ich kann lange, komplexe Sachtexte und literarische Texte verstehen und Stilunterschiede wahrnehmen. Ich kann Fachartikel und längere technische Anleitungen verstehen, auch wenn sie nicht in meinem Fachgebiet liegen.
B2	Ich kann längere Redebeiträge und Vorträge verstehen und auch komplexer Argumentation folgen, wenn mir das Thema einigermaßen vertraut ist. Ich kann am Fernsehen die meisten Nachrichtensendungen und aktuellen Reportagen verstehen. Ich kann die meisten Spielfilme verstehen, sofern Standardsprache gesprochen wird.	Ich kann Artikel und Berichte über Probleme der Gegenwart lesen und verstehen, in denen die Schreibenden eine bestimmte Haltung oder einen bestimmten Standpunkt vertreten. Ich kann zeitgenössische literarische Prosatexte verstehen.
B1	Ich kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. Ich kann vielen Radio- oder Fernsehsendungen über aktuelle Ereignisse und über Themen aus meinem Berufs- oder Interessengebiet die Hauptinformation entnehmen, wenn relativ langsam und deutlich gesprochen wird.	Ich kann Texte verstehen, in denen vor allem sehr gebräuchliche Alltags? oder Berufssprache vorkommt. Ich kann private Briefe verstehen, in denen von Ereignissen, Gefühlen und Wünschen berichtet wird.
A2	Ich kann einzelne Sätze und die gebräuchlichsten Wörter verstehen, wenn es um für mich wichtige Dinge geht (z. B. sehr einfache Informationen zur Person und zur Familie, Einkaufen, Arbeit, nähere Umgebung). Ich verstehe das Wesentliche von kurzen, klaren und einfachen Mitteilungen und Durchsagen.	Ich kann ganz kurze, einfache Texte lesen. Ich kann in einfachen Alltagstexten (z. B. Anzeigen, Prospekten, Speisekarten oder Fahrplänen) konkrete, vorhersehbare Informationen auffinden und ich kann kurze, einfache persönliche Briefe verstehen.
A1	Ich kann vertraute Wörter und ganz einfache Sätze verstehen, die sich auf mich selbst, meine Familie oder auf konkrete Dinge um mich herum beziehen, vorausgesetzt es wird langsam und deutlich gesprochen.	Ich kann einzelne vertraute Namen, Wörter und ganz einfache Sätze verstehen, z. B. auf Schildern, Plakaten oder in Katalogen.

4.1.Lesekompetenz testen

Grotjahn (2000c) erweitert die oben eingeführten Lesestrategien, und beschreibt sie abhängig von den Zielen des Lesers:

- **suchendes Lesen/ *search reading***: selektive Lokalisierung von Informationen zu festen Fragestellungen;
- **identifizierendes Lesen/ *scanning***: Suche nach einer bestimmten Zeichenkette wie Namen oder Zahlen;
- **orientierendes Lesen/ kursorisches Lesen/ *skimming***: Ziel ist ein Überblick über den Textinhalt; das Lesen folgt Textaufbau und versucht wesentliches zu erfassen;
- **intensives Lesen/ detailliertes Lesen**: sorgfältiges Lesen mit dem Ziel eines globalen Verständnisses und des Erfassens von Details;
- **argumentatives Lesen/ *responsive reading***: Auseinandersetzung mit dem Textinhalt auf der Basis des eigenen Weltwissens;
- **extensives lesen**: relativ schnelle Lektüre großer Textmengen, in der Regel zum Vergnügen. (Grotjahn 2000c: 23f.)

Diese Lesestile korrespondieren mit den gelesenen **Textsorten**, aber auch mit typischen Aufgabenstellungen, so werden Telefonbücher wohl kaum in- und extensiv gelesen (Grotjahn 2000c: 25). Geschlossene Aufgaben zielen häufig auf das Detailverständnis eines Testes und nicht auf ein globales Verständnis. Offene Aufgaben aber benötigen häufig ein intensives oder gar argumentatives Lesen auf höherem Kompetenzniveau.
Die **Testinstruktionen** für einen Lesetest sollten eindeutig und dem Niveau angemessen sein. Sie sollten für Aufgabentypen einheitlich sein, um dem Testteilnehmer das Verständnis zu erleichtern und mit folgenden Hinweisen versehen werden (Grotjahn 2000c: 72ff.):

- Ort der Textvorlage (Seitenzahl),
- Umfang der Antwort,
- Welche Aufgaben/ wie viele Aufgaben gelöst werden müssen,
- Anzahl der richtigen Antworten in Multiple-Choice,
- Zulassung einer mehrmaligen Verwendung von Wahlmöglichkeiten in Zuordnungsaufgaben,
- Arbeitszeit.

Zur Vorhersage der **Aufgabenschwierigkeit** gibt es mehrere Ansätze (Grotjahn 2000c: 88ff.): **Textzentrierte Ansätze** bestimmen die Schwierigkeit als textimmanente Eigenschaft, die häufig bestimmt durch Expertenratings in vier Dimensionen auf einer fünfstufigen bestimmt bipolaren Skala wird. Die Dimensionen kann man sich hierbei wie folgt vorstellen:

1	Einfachheit		Kompliziertheit
2	Gliederung/ Ordnung		Zusammenhang-losigkeit
3	Kürze/ Prägnanz	**Schwierigkeit**	Weitschweifigkeit
4	Zusätzliche Stimulanz		fehlende zusätzliche Stimulanz

Abbildung 14: Die Dimensionen der Schwierigkeit von Texten

Adressatenzentrierte Ansätze auf der anderen Seite bestimmen die Aufgabenschwierigkeit empirisch anhand der Lösungen von Testkandidaten. Grotjahn (2000c) fasst die Ergebnisse von Pollitt (1986: 55 ff) über Gründe für die Schwierigkeit von Testfragen so zusammen

„1. Die Formulierung der Frage ist potentiell mehrdeutig.
2. Der für die Beantwortung der Frage relevante Textteil enthält Inhaltswörter, die Einstellungen und Emotionen ausdrücken und funktionale Details überdecken.
3. Die für eine Antwort benötigte Information ist über den Text verteilt.
4. Eine Frage besteht aus zwei oder mehr Teilen, die in komplexer Weise miteinander verbunden sind.
5. Die Formulierung der Antwort verlangt mehr als eine wörtliche oder leicht modifizierte Wiedergabe des Textes.
6. Fehler bei vorangehenden Antworten können zu Fehlern bei folgenden Antworten führen.
7. Die Satzstruktur des Textes ist komplex.“ Grotjahn (2000c: 96)

Aus einer Untersuchung von Buck et al. (1997) folgt unter anderem für die Schwierigkeit von Multiple-Choice-Tests des Leseverstehens, dass leichte Items eine hohe lexikalische Übereinstimmung zwischen der korrekten Lösungsoption und der Textinformation haben und die notwendige Textstelle ist leicht zu finden oder offensichtlich ist. Bei schweren Items indes stimmt zum Teil die Reihenfolge der Items nicht mit der Reihenfolge des Textes überein, sie verlangen eine Inferenz aus zwei Sätzen oder die korrekte Option hat keine lexikalische Überlappung. Außerdem kann es sein, dass die zur Lösung notwendige Information über den gesamten Text verteilt ist. Buck et al. (1997: 441ff.)

4.2.Hörverstehen testen

Das Hörverstehen ist eine rezeptive Fertigkeit, die mit einer spezifischen Schwierigkeit verbunden ist: Hörtexte sind kontinuierliche akustische Signalen in Echtzeit. Der Hörer muss daher als Teil des Verstehensprozesses einen kontinuierlichen akustischen Strom segmentieren, was, je nach regionaler Färbung und entsprechenden Koartikulationen und Reduktionen mehr oder weniger schwierig sein kann. Außerdem hat der Hörer wenig Kontrolle über die Geschwindigkeit des Textes, er kann die Rezeption nicht beliebig oft wiederholen und damit sein Verständnis korrigieren. Außerdem ist ein Hörverstehenstest störanfällig durch die eventuelle mangelnde Qualität der Aufnahme oder störende Nebengeräusche. Das Hörverstehen von Tonaufnahmen ist immer schwieriger als das Hörverstehen im Alltag, da hier das Kontextwissen der Hörer oft beschränkt ist und die visuelle Komponente fehlt. Deshalb sind für solche Übungen und Tests Videoaufnahmen zu empfehlen, da Bilder, Gesten etc. als Kontext zum Verständnis beitragen.
Als sinnvoll hat sich bei Hörverstehenstests erwiesen, den Text erst als ganzen hören zu lassen, und erst dann die Aufgaben zu lesen um beim zweiten Hören gezielter nach Informationen suchen zu können. Dadurch können sich die Prüflinge an akustische und sprachliche Besonderheiten gewöhnen. Außerdem bauen sie „Verstehensinseln“ auf (Solmecke 2000: 68), die eine Orientierung im Text und ein erleichtertes Verständnis bei der Wiederholung ermöglichen.
Nach Solmecke (2000) lassen sich akustischen Schwierigkeiten des Textes als Zusammenspiel zwischen technischer Akustik und Hintergrundgeräuschen verstehen, es zählen aber auch überlappende Sprecherwechsel und Unterbrechungen dazu. Die linguistischen Schwierigkeiten eines Hörtestes schlüsselt er wie folgt auf:

Textschwierigkeiten						
	← eher leicht					eher schwierig →
		1	2	3	4	
Wortschatz	alltäglich					spezialisiert
Steilheitsgrad	niedrig					hoch
Erschließbarkeit	problemlos					schwierig
Unübliche Wortkombinationen	keine					häufig
Ambiguitäten	keine					häufig
Feststehende Redewendungen	erschließbar					nicht erschließbar
Satzlänge	kurz					lang
Satzkomplexität	gering					hoch
Grammatische Redundanz	hoch					niedrig
Versprecher/Pausenfüller	keine					häufig
Andere Normabweichungen: Lexik	keine					häufig
Andere Normabweichungen: Grammatik	keine					häufig
Kommunikative Konventionen	neutral					kulturspezifisch
Diskursart	informell					formell
Textsorte	neutral					kulturspezifisch
Verhältnis von Textart und Textinhalt	üblich					unüblich
Textstruktur	klar					verschachtelt
Explizite Gliederungssignale	häufig					keine
Textkohärenz/ -kohäsion	hoch					niedrig
Textfunktionen implizit	nie					häufig

Abbildung 15: Linguistische Schwierigkeit eines Hörtextes (Solmecke 2000: 64)

Die inhaltlichen Schwierigkeiten eines Textes ergeben sich aus der Explizitheit und Direktheit auf verschiedenen Ebenen (Information, Intention etc.) und aus dem Sachwissen der Prüflinge. Hinzu kommt die Aufgabenschwierigkeit, die einerseits, wie wir bereits oben gesehen haben, abhängig ist vom Aufgabentyp, andererseits natürlich von der sprachlichen Verständlichkeit der Aufgabenstellung sowie ihrer Eindeutigkeit beeinflusst wird. Bei offenen Aufgaben ist auch die Fähigkeit der Testperson, die geforderte Sprachhandlung auszuführen, zu beachten. Auch spielt die Textangemessenheit der Aufgabe eine Rolle, denn ein unterhaltender Text ist beispielsweise häufig nicht explizit genug für eine detaillierte Informationsentnahme.

Über die von Bolton (1996: 22) genannte Hörstile hinaus unterscheiden Rost (1990) und (2000b) **Hörerrollen** anhand der Kooperation von Sprecher und Hörer (Rost 1990: 5fff., Grotjahn (2000b: 7). Auch diese führen zu unterschiedlichen Anforderungen: **Gesprächsteilnehmer**, die gleiche Rederechte haben, können ihr Verständnis anhand der Reaktionen von anderen Teilnehmern prüfen. Dies ist nicht der Fall für **Adressaten**, die vom Sprecher angesprochen werden, aber nur eingeschränkt antworten dürfen (wie in einer Vorlesung). Auch der **Zuhörer** einer Einwegkommunikation, der zwar intendierter Adressat ist, aber nicht Antworten kann und soll, kann sein Verständnis nicht durch Nachfragen überprüfen. Am schwierigsten ist das Hörverstehen jedoch für einen **Mithörer**, der nicht Adressat ist: Ihm fehlt das Situationswissen und Sachwissen der Sprecher. Dennoch ist dies leider die typische Rolle in Hörverstehensübungen und -tests.

Von der Hörerrolle abhängig ist auch die **Verstehensabsicht** eines Hörers. Auch hier ist ein Hörverstehenstest natürlich nur eingeschränkt authentisch, denn oft genug ist in Tests die Verstehensabsicht unspezifisch, daher versuchen Lernende in testähnlichen Übungen häufig, den Text wortwörtlichen zu verstehen, was dem Verstehen hinderlich ist. Die

Verstehensabsicht wird lediglich durch die Testaufgabe konkretisiert und normiert, allerdings ist nicht davon auszugehen, dass eine solche fremdbestimmte Konkretisierung auch dazu führt, dass der einzelne Testteilnehmer diese Verstehensabsicht tatsächlich erkennt und übernimmt (Solmecke 2000: 60). All diese Probleme eines Hörverstehenstests können eigentlich nur dadurch gemindert werden, dass sie im Unterricht thematisiert werden, indem man mit den Lernenden in einen Dialog über ihre Probleme beim Verstehen eintritt, sie darauf hinweist, dass sie der künstlichen Situation geschuldet sind, und gemeinsam mit ihnen Strategien zur Überwindung der Probleme entwickelt.

Trotz oder vielleicht auch wegen der Probleme der mangelnden **Authentizität** der Testsituation wird in der Literatur die Authentizität der verwendeten Hörtexten als Qualitätskriterium betont und verlangt, dass, wenn möglich, originale Rundfunk- und Fernsehsendungen eingesetzt werden. Bolton (1996) weist allerdings darauf hin, dass dies in der Grundstufe problematisch ist. Als mögliche Alternativen schlägt sie Bahnhofsdurchsagen/ Flughafenansagen, Staumeldungen und einfache Schulsendungen vor. Es sei natürlich auch möglich, reale Alltagsdialoge selbst aufzunehmen. Hierbei arbeite man am besten mit Kollegen zusammen, um eine einseitige Fokussierung der Lernenden auf die eigene Stimme und Aussprache zu vermeiden und einen gewissen Fremdheitsschock zu simulieren. (Bolton 1996: 46)

Da die Aufgabentypen im Wesentlichen denen im Leseverstehen (Abschnitt 4.1) entsprechen, sollen sie hier nicht separat besprochen werden.

5. Produktive Fertigkeiten Testen: Schreiben und Sprechen

Neben dem oben besprochenen Problem der Bewertung von offenen Aufgabentypen, wie sie beim Prüfen der produktiven Fertigkeiten hauptsächlich Verwendung finden, spielt eine andere Problemfrage eine große Rolle: Was genau soll in einem Test der produktiven Fähigkeiten getestet werden? Im Unterricht sind Sprech- und Schreibaufgaben schließlich oft kein Selbstzweck und dienen nicht dem Erwerb der übergeordneten Sprech- und Schreibkompetenz, sondern sie dienen der Festigung des Wortschatzes und der grammatischen Strukturen. Damit stellen Sprechen und Schreiben hier eine **Mittlerfertigkeit** dar.

Auf der anderen Seite stellen Sprech- und Schreibprüfungen integrierte Performanztests dar, denn sie prüfen nicht nur die kommunikativen Fertigkeiten an sich, sondern werden von vielen Lehrern und Prüfern auch als Prüfung der Lexik und Grammatik verstanden, die in Kompetenztests getrennt geprüft wird. Da aber reine Kompetenztests kritisch betrachtet werden müssen – schließlich testen sie nicht die sprachliche Alltagstauglichkeit der Kandidaten – ist eine integrierte Prüfung der Teilkompetenzen sinnvoll, bereitet aber andererseits Schwierigkeiten bei der Bewertung. Soll beispielsweise ein Kandidat, der keine Fehler macht, aber nur einfachste Konstruktionen verwendet, genauso gut bewertet werden wie ein Kandidat, der komplexe Konstruktionen verwendet, aber mehr Fehler macht? Auch wenn man dies übereinstimmend ablehnt, bleibt immer noch die schwierige Frage zu beantworten, wie genau man die Komplexität der Formulierungen bestimmt, um diese dann gegen die Fehlerzahl aufzurechnen.

5.1.Schreibfertigkeit testen

(1) Folgende Teilfertigkeiten, die im Unterricht vermittelt werden müssen und in Prüfungen getestet werden sollten, spielen beim Schreiben eine Rolle:

- Orthografie,
- Vokabular,
- Satzgrammatik (Satzstellung und -typen, Kasussystem, Verbvalenz etc.),
- Textgrammatik (Vertextungsmittel wie Verweise (Pronomina, Deiktika etc.) und Konnektoren; Makro- und Superstrukturen von Texten wie Einleitung, Hauptteil, Schluss oder Anreden),
- Textsorten (z.B. Brief; Textaufbau).

Diese Teilfertigkeiten unterscheiden sich für die verschiedenen **Schreibformen** wie Informieren, Argumentieren, Appellieren, Untersuchen und Gestalten. So verlangt das Erzählen beispielsweise die Beherrschung der Vergangenheitsformen und Zeitadverbien, während das Argumentieren u.a. den Gebrauch des Konjunktivs und eine bestimmte Textstruktur verlangen kann.
In der Grundstufe ist es durchaus noch möglich, die Teilfertigkeiten getrennt zu prüfen, z.B. kann man als Aufgaben zur Vertextung unverbundene Hauptsätze zu einem Text umformulieren lassen oder mit einem Lückentexte die Beherrschung der Pronomina prüfen. Beim integrierten Testen der Schreibfertigkeit durch das Verfassen zusammenhängender Texte ist zu beachten, dass die Aufgabenstellung im Hinblick auf Thema, die erwartete Textsorte und die Textlänge eindeutig ist, und die Textsorte Gegenstand des Unterrichts gewesen sein muss. Außerdem soll die Testaufgabe eine Vergleichbarkeit der Testergebnisse erlauben (vgl. Bolton 1996: 76f.).

(2) Eine Möglichkeit, eine bessere Vergleichbarkeit herzustellen und damit die Reliabilität einer Schreibprüfung zu erhöhen, ist die Verwendung **vorstrukturierter Schreibaufgaben.** Je enger die Aufgabenstellung ausfällt, desto vergleichbarer sind die Ergebnisse, wodurch die Reliabilität des Testes zunimmt (Dlaska/Krekeler 2009: 114). Hierzu können die Inhalte und das sprachliche Material je nach Sprachstand der Schüler durch **Leitpunkte** vorgegeben werden. Diese strukturieren den Text Schritt für Schritt und erlauben dem Kandidaten, sich auf die sprachliche bzw. textlinguistische Arbeit am Produkt zu konzentrieren (Bolton 1996: 82f.).

Beispielaufgabe

> Schreiben Sie einen Brief an einen deutschen Brieffreund, in dem Sie über Ihre Reise in den Sommerferien berichten. Schreiben Sie etwas über
>
> - ❖ Wann, wie, mit wem und wohin Sie gereist sind;
> - ❖ Wo Sie gewohnt haben (ein Hotel, ein Ferienhaus, was gab es zu essen);
> - ❖ Was sie besonders schön fanden, z.B. ein Museum, den Strand, ein Restaurant.
>
> Vergessen Sie auch nicht Datum, Anrede, Gruß und Unterschrift!

Statt mit Leitpunkten viel Struktur und sprachliches Material vorzugeben, lässt auf sich für Lerner der Mittelstufe der erwartete Text durch einen **Vorspann** in eine kommunikative Situation einordnen:

Beispielaufgabe

> Sie haben im letzten Sommer bei einem Sprachkurs in Deutschland mit dem Deutschen Medizinstudenten Max zusammengewohnt und stehen seitdem in Briefkontakt. Schreiben Sie ihm nun einen Brief, in dem Sie Ihren Sommerurlaub beschreiben.

Ein weiteres Stimuluselement, welches die Schreibaufgabe vorstrukturiert, kann ein Text zum Thema sein. Dies ist vor allem im akademischen Schreiben relevant, da dort der Verweis auf andere Texte Teil der Textsortenkompetenz ist. Außerdem geben Textvorgaben das erforderliche Hintergrundwissen vor und verhindern dadurch Schreibblockaden. Hierbei kann die Vorgabe mehrerer kleinerer Texte sogar erfolgreicher sein (vgl. Smith et al. 1985). Zu beachten ist allerdings, dass hierdurch die Lesekompetenz Teil der Bewertung wird, da schlechte Leser bestraft werden (Weigle 2002: 95).
Testpersonen eine **Auswahl von Aufgaben** anzubieten senkt zwar die Reliabilität des Tests, da die Texte nicht vergleichbar sind und eine exakt gleiche Schwierigkeit der Aufgaben schwer zu erreichen ist. Es kann aber unter Umständen die Motivation wesentlich erhöhen, eine mögliche Prüfungsangst verringert und die Abhängigkeit des Testergebnisses vom Hintergrundwissen verringern.

(3) Als **Themenbereiche** in Schreibprüfungen eignen sich Themen der persönlichen Erfahrung oder Themen des Allgemeinwissens (Weigle 2002: 92f.). Themen aus der persönlichen Erfahrung haben den Vorteil, dass die Prüflinge sich mit ihnen gut identifizieren können und leicht Material in ihren Erfahrungen finden, da kein spezielles Wissen verlangt wird. Außerdem erzeugt ein breiteres Themenfeld größeres Interesse beim Prüfer, was für eine gleichmäßige konzentrierte Korrektur durchaus hilfreich ist. Allerdings können solche Themen für Studenten aus Kulturen, in denen der schriftliche Selbstausdruck nicht wertgeschätzt wird, problematisch sein. Der Selbstausdruck in diesen Texten kann außerdem den Fokus des Prüflings von der eigentlichen Textproduktion ablenken, da er emotional stark beteiligt ist. Bei Themen des Allgemeinwissens ist zu beachten, dass möglicherweise fehlendes Hintergrundwissen das Prüfergebnis beeinflussen kann. Dies ist allerdings hauptsächlich bei curriculumsunabhängigen Sprachstandsprüfungen problematisch, da bei Prüfungen, die sich auf einen bestimmten Sprachkurs beziehen, die Aufgaben sich auf im Unterricht behandelte Themen beziehen und dieses Wissen zur Voraussetzung machen können. Hier ist dann die Entscheidung zu fällen, ob z.B. Landeskundewissen mit geprüft werden soll – dann muss es unterrichtet worden sein, oder ob es in der Prüfung nur um Schreibkompetenz geht.

(4) Da bei mündlichen wie schriftlichen Prüfungen die R-Komponente ein im Grunde freier Text ist, kann die Bewertung nicht nach einem simplen „richtig/falsch“ oder „gelöst/ nicht gelöst“ Muster erfolgen, sondern muss **Methoden der Textbewertung** heranziehen, die für mündliche und schriftliche Prüfungen in abgewandelter Form gelten.
Lange Zeit war der Lehrer oder Prüfer das Maß der Dinge, da Texte durch eine **Bewertung nach Gesamteindruck** beurteilt wurden. Hierbei werden mehr oder weniger transparente interne Kriterien angewandt, bei schriftlichen Arbeiten vergleichen die Prüfer oft die vorliegenden Ergebnisse und bringen sie in eine Reihenfolge. Dies wird heutzutage in der Regel als kriterienlos und intransparent verurteilt, allerdings weisen Dlaska und Krekeler (2009) darauf hin, dass diese Bewertungsmethode auch Chancen bietet, einem Text offen zu begegnen und die Reliabilität erhöht werden kann, wenn Texte im Vergleich zu Texten aus

der Lerngruppe bewertet werden, indem man die Texte nach einer ersten Sichtung gruppiert (Dlaska/Krekeler 2009: 106).

Eine häufig verwendete Methode, die Bewertung von Texte zu verobjektivieren, stellt die **fehlerorientierte Bewertung** dar, bei der in der Regel der Fehlerquotient[4] zu Grunde gelegt wird, um eine Relation zwischen Fehlerzahl und Textlänge zu erreichen. Hierbei bleibt allerdings die Qualität der Fehler unberücksichtigt, es sei denn man definiert ganze und halbe Fehler und führt hierdurch eine Gewichtung der Fehler ein. Problematisch bleibt dann immer noch die Definition der Schwellenwerte für Quotienten, d.h. der Umrechnung des Fehlerquotienten in Noten(punkte). Außerdem erfasst diese Art der Bewertung nicht die Bewertung der eigentlichen Qualitäten des Textes wie Satzlänge/ Struktur/ Vokabular und sollte daher nicht allein bestimmend, sondern nur ein Teil der Bewertungskriterien sein.

Eine weitaus transparentere Methode ist die **ganzheitliche Bewertung** (*holistic scoring*), bei der das Produkt nach einer Gesamtdefinition für die entsprechende Leistung bewertet wird (Hughes 2003). In dem Kriterienkatalog werden Ergebnisklassen beschrieben, ohne die Kriterien zu gewichten. Dies kann problematisch sein, da die Teilfertigkeiten manchmal ungleichmäßig ausgebildet sind. Als Beispiel ist in

Abbildung 16 das Bewertungsraster für den schriftlichen Teil des TOEFL abgebildet. Hierbei ist allerdings zu beachten, dass gerade beim TOEFL die Bewerterobjektivität noch durch andere Methoden abgesichert wird. So unterlaufen die Bewerter ein intensives Training und müssen regelmäßig bereits vorbewertete Texte bewerten, um sicher zu stellen, dass alle die Kriterien des Katalogs gleichmäßig anwenden.

Abbildung 16: TOEFL writing scoring guide **(Weigle 2002: 144)**

6 An essay at this level

- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in use of language
- demonstrates syntactic variety and appropriate word choice though it may have occasional errors

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level

- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

3 An essay at this level may reveal **one** or more of the following weaknesses:

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations

[4] $Fehlerquotient = \frac{Fehlerzahl \times 100}{Gesamtzahl\ der\ geschriebenen\ Wörter}$

- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by one or more of the following weaknesses:

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level

- may be incoherent
- may be undeveloped
- may contain severe and persistent writing errors

0 A paper is rated 0 if it contains no response, merely copies the topic, is off- topic, is written in a foreign language, or consists of only keystroke characters.

Bei der **analytischen Bewertung** findet ein differenzierter und gewichteter Kriterienkatalog Anwendung. Diese Methode ist verlässlicher als die ganzheitliche Bewertung, da sie die Teilkompetenzen getrennt aufschlüsselt und in ihrer Gewichtung bewertet. Damit wird der Entscheidungsspielraum des einzelnen Prüfers eingeschränkt. In Abbildung 17 findet sich ein Kriterienkatalog, der für ein Sprachenzentrum an einer britischen Universität entwickelt wurde. Wie bereits erläutert, wird im britischen System in Prozentpunkten bewertet, wobei die Grenze zum „Ausgezeichnet“ bei 70 Prozent liegt und 80 Prozent sehr sehr selten überschritten wird. Daher sind die Teilpunkte in diesem Raster ebenfalls so definiert, dass nur in außergewöhnlichen Fällen die Teilnote 80 Prozent überschritten wird. An diesem Beispiel lassen sich aber die generellen Charakteristika und Probleme analytischer Kriterienkataloge erkennen. Unser Beispielkatalog ist, wie zu erwarten, definiert für die verschiedenen Teilkompetenzen. Andererseits ist dieser Kriterienkatalog definiert für verschiedene Sprachen und alle Lernstufen, wodurch das Urteil über die Stufen- und Sprachenangemessenheit weiterhin in der Hand des Prüfers liegt. Dies könnte natürlich durch präzisere Definitionen, beispielsweise unter Angabe von erwarteten grammatischen Strukturen oder Textsorten verbessert werden. Allerdings sind analytische Raster eben niemals vollständig objektiv und reliabel, denn eine absolut präzise Definition eines bestimmten Leistungsspektrums wäre letztlich unendlich lang und detailliert und kaum zu handhaben. Daher sind solche Kataloge immer ein Kompromiss zwischen Komplexität und Handhabbarkeit.
Beide Arten der Kriterienkataloge – analytische wie holistische – müssen allerdings weitere Problemfälle antizipieren: Sie müssen regeln, wie Antworten, die das Thema der Frage nicht behandeln, unvollständige sowie auswendig gelernte Antworten zu bewerten sind (Weigle 2002: 131ff.). Bei Antworten, die für die Frage nicht relevant sind, ist zu entscheiden, inwieweit die inhaltliche Beantwortung relevant ist. Ein Argument gegen eine Bestrafung inhaltlicher Verfehlungen ist, das in einem Test der Schreibkompetenz die Lesekompetenz, die für das Erfassen der Aufgabe nötig ist, eigentlich nicht getestet werden soll. Dieses Argument ist allerdings nur gültig bei hoch standardisierten Tests, da gerade bei informellen Tests oder bei schulischen Sprachprüfungen der Inhalt Teil der Prüfung ist. Zusätzlich gilt, dass die erwartete Aufgabenspezifik auswendig gelernten Antworten vorbeugen, die nicht den eigentlichen Leistungsstand wiedergeben. Eine Möglichkeit mit diesen wie auch dem Problem der unvollständigen Antworten umzugehen ist, die Punktevergabe für die Sprache und den Inhalt so zu koppeln, dass die Bewertung für die Sprache nicht besser sein darf als die Bewertung des Inhalts.

Abbildung 17: Bewertungsraster für eine schriftliche Prüfung in einer Europäischen Fremdsprache

Points out of 10	***Organization of Ideas***
8(up to 10)	Outstanding organisation and control of material. Outstanding degree of logic and coherence throughout. Exceptionally skilful handling of material.
7	Excellent organisation and control of material. A high degree of logic and coherence throughout. Very skilful handling of material.
6	Material very well marshalled and developed within carefully planned framework. Logical sequence of ideas. Skilfully controlled throughout.
5	Good organisation and sequencing of ideas. Lacks coherence in places. Occasionally repetitive.
4	Satisfactory organisation. Development patchy and/or lacking ambition. Often repetitious and/or rambling.
below 4	Disorganised and lacking coherence throughout OR not enough to evaluate
Points out of 20	***Vocabulary***
16 (up to 20)	Outstanding command of lexis and structures.
14-15	Excellent range of lexis and good variety of structures with only very little limitation.
12-13	Good range of lexis with some examples of more complex structures.
10-11	Adequate range of lexis; limited range of structures.
8-9	Very basic lexis.
below 8	Only minimal command of structure OR not enough to evaluate
Points out of 50	***Language Use (Grammar/ Pragmatics)***
40 (up to 50)	Effective complex constructions, virtually no errors of agreement, tense, number, word order/function, articles, pronouns, prepositions

35-39	Effective complex constructions, very few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions.
30-34	Effective constructions but some problems in complex constructions, some errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
25-29	Effective but simple constructions; problems in complex constructions; several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
20-24	Major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions; meaning confused or obscured
below 20	virtually no mastery of sentence construction rules; dominated by errors; does not communicate OR not enough to evaluate
Points out of 20	**Mechanics**
16 (up to 20)	Demonstrates mastery of conventions, virtually no errors of spelling, punctuation, capitalization, paragraphing.
14-15	Demonstrates mastery of conventions, few errors of spelling, punctuation, capitalization, paragraphing.
12-13	Occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured.
10-11	Occasional errors of spelling, punctuation, capitalization, paragraphing; poor handwriting; meaning confused or obscured
8-9	Frequent errors of spelling, punctuation, capitalization, paragraphing; poor handwriting, meaning confused or obscured
below 8	No mastery of conventions; dominated by errors of spelling; punctuation, capitalization, paragraphing; handwriting illegible OR not enough to evaluate

Result: ……/100

Weigle (2002) stellt eine weitere Methode zur Bewertung schriftlicher Arbeiten vor: „**primary trait scales**“. Diese aufgabenspezifischen Bewertungshorizonte, wie sie ähnlich im deutschen Abitur verwendet werden, sind formale Dokumente für die einzelnen spezifischen Essayfragen, die folgende Informationen enthalten müssen:

- die Schreibaufgabe;
- eine Definition der primären rhetorischen Charakterzüge (z.B. argumentativer Aufsatz, Glückwunschkarte);
- eine Hypothese über die erwarteten Leistungen;
- eine Darstellung des Zusammenhangs zwischen primärem rhetorischen Charakter und der Aufgabe;
- Beispiellösungen für jedes Bewertungsniveau;
- Erläuterung zur Bewertung der Beispiellösungen. (Weigle 2002: 110)

Die Autorin weist allerdings selbst darauf hin, dass ein solcher Bewertungshorizont sehr aufwändig ist, speziell wenn auch noch empirisch vorgetestet werden soll, weshalb er in der Praxis in diesem Umfang kaum Anwendung findet.

	Schreiben
C2	Ich kann klar, flüssig und stilistisch dem jeweiligen Zweck angemessen schreiben. Ich kann anspruchsvolle Briefe und komplexe Berichte oder Artikel verfassen, die einen Sachverhalt gut strukturiert darstellen und so dem Leser helfen, wichtige Punkte zu erkennen und sich diese zu merken. Ich kann Fachtexte und literarische Werke schriftlich zusammenfassen und besprechen.
C1	Ich kann über eine Vielzahl von Themen, die mich interessieren, klare und detaillierte Texte schreiben. Ich kann in einem Aufsatz oder Bericht Informationen wiedergeben oder Argumente und Gegenargumente für oder gegen einen bestimmten Standpunkt darlegen. Ich kann Briefe schreiben und darin die persönliche Bedeutung von Ereignissen und Erfahrungen deutlich machen.
B2	Ich kann über Themen, die mir vertraut sind oder mich persönlich interessieren, einfache zusammenhängende Texte schreiben. Ich kann persönliche Briefe schreiben und darin von Erfahrungen und Eindrücken berichten.
B1	Ich kann über Themen, die mir vertraut sind oder mich persönlich interessieren, einfache zusammenhängende Texte schreiben. Ich kann persönliche Briefe schreiben und darin von Erfahrungen und Eindrücken berichten.
A2	Ich kann kurze, einfache Notizen und Mitteilungen schreiben. Ich kann einen ganz einfachen persönlichen Brief schreiben, z. B. um mich für etwas zu bedanken.
A1	Ich kann eine kurze einfache Postkarte schreiben, z. B. Feriengrüße. Ich kann auf Formularen, z. B. in Hotels, Namen, Adresse, Nationalität usw. eintragen.

Abbildung 18: GeR Deskriptoren für eine Selbstevaluation der Schreibkompetenz (Trim et al. 2009: 36)

5.2.Sprechfertigkeit testen

(1) Auch bei mündlichen Prüfungen in den Fremdsprachen stellt sich wieder die Frage, was eigentlich geprüft werden soll, denn auch die Fertigkeit Sprechen enthält natürlich Teilfertigkeiten wie Aussprache und Grammatik. Bereits bei der Bewertung der Grammatik ergeben sich Unterschiede zur Bewertung schriftlicher Leistungen, denn hier geht es um die Grammatik gesprochener Sprache, die auch bei Muttersprachlern von der Schriftgrammatik erheblich abweichen kann, weshalb eine Bewertung hier wesentlich toleranter sein sollte. Der Kern der Fertigkeit ist aber das sprachliche Handeln, bei dem Bygate (1987) folgendes unterscheidet:

Faktisches Sprechen	Bewertendes Sprechen
Beschreiben	**Erklären**
Erzählen	**Rechtfertigen**
Anleiten	**Vorhersagen**
Vergleichen	**Entscheiden**

Abbildung 19: Sprechhandlungen (Bygate 1987)

Diese Arten des Sprechens, die den Sprechhandlungstypen der Sprechakttheorie verwandt sind, können als Makrofunktionen des Sprechens verstanden werden und sind als solche in den GeR eingegangen (Luoma 2004: 32f.). Als Mikrofunktionen werden hingegen Kategorien verstanden, die einzelne Redebeiträge eines Sprechers in einer Interaktion betreffen:

„1.1 Sachinformationen mitteilen und erfragen

- identifizieren
- berichten
- berichtigen
- fragen
- antworten

1.2 Einstellungen zum Ausdruck bringen und erfragen in Bezug auf:

- Fakten (Zustimmung/Ablehnung)
- Wissen (Wissen/Unwissen, Erinnern/Vergessen, Wahrscheinlichkeit/Sicherheit)
- Modalität (Verpflichtung, Notwendigkeit, Fähigkeit, Erlaubnis)
- Willensbekundungen (Wünsche, Verlangen, Absichten, Präferenzen)
- Gefühle (Freude/Missfallen, Vorlieben/Abneigungen, Zufriedenheit, Interesse, Überraschung, Hoffnung, Enttäuschung, Angst, Sorge, Dankbarkeit)
- Moralisches Verhalten (Entschuldigungen, Billigungen, Bedauern, Mitgefühl)

1.3 Überreden, Überzeugen

- Vorschläge, Bitte, Warnung, Rat, Ermutigung, um Hilfe bitten, Einladungen, Angebote

1.4 Soziale Routinen

- Aufmerksamkeit erregen, jemanden anreden und (be)grüßen, jemanden vorstellen, einen Trinkspruch ausbringen, sich verabschieden

1.5 Diskursstrukturierung

- (28 Mikrofunktionen, Eröffnen, Sprecherwechsel, Abschließen usw.)

1.6 Reparaturen, Selbstkorrektur

- (16 Mikrofunktionen)

(Trim et al. 2009: 126)

Aus diesen kommunikativen Funktionen sollten sich einerseits die Prüfungsaufgaben ergeben, andererseits können mit ihrer Hilfe ähnliche Kriterienkataloge wie die in Kapitel 5.1 vorgestellten holistischen oder analytischen Systeme erarbeitet werden.

(2) Am problematischsten scheint aber immer die Frage zu sein, in welchem **Rahmen** geprüft wird. Hierbei spielen in der Regel drei Kategorien eine Rolle

- **Kandidaten**: einer, zwei, Gruppe
- **Prüfer**:
  - eine Person vereint Gesprächsleiter und Prüfer
  - ein Gesprächsleiter und ein Prüfer sind gemeinsam in einem Raum
  - ein Gesprächsleiter nimmt das Gespräch elektronisch auf, es wird von einem nicht anwesenden Prüfer („blind“) bewertet
  - eine elektronische Aufnahme stellt Fragen, die Antworten werden elektronisch aufgezeichnet und von einem nicht in der Prüfung anwesenden Prüfer bewertet
- gibt es eine **Vorbereitungszeit** – und was wird vorbereitet?

Als Zeitrahmen einer mündliche Prüfung ist es sinnvoll zwischen 15 und 30 Minuten einzuplanen – Hughes (2003) geht davon aus, dass sich in weniger als 15 Minuten sich kaum sinnvolle Informationen über die Kompetenz des Kandidaten sammeln lassen. Um eine Vergleichbarkeit herzustellen, sollte wenigstens ein Grobskript mit den wesentlichen Zielen, Impulsen (Fragen/ Aufforderungen) und erwarteten Antworten sowie den Zeiten für die Aufgaben erstellt werden. Hilfreich ist, mehrere kleine verschiedene Aufgaben zu stellen, um sowohl verschiedene Diskursarten zu prüfen als auch dem Prüfling viele Chancen für Neuanfänge zu geben, damit nicht vielleicht gerade das eine für den Prüfling uninteressante Thema das Ergebnis der Prüfung bestimmt. (Hughes 2003: 124)

Die klassische mündliche Prüfung in den Fremdsprachen stellte – oft vor dem Hintergrund beschränkter Ressourcen – eine Interviewsituation mit **einem** Prüfer dar. Hierbei ist vor allem die **Fragetechnik** des Prüfers entscheidend, denn der Einfluss des Prüfers auf die Leistungen einer mündlichen Prüfung in den Fremdsprachen ist gut belegt (z.B. Brown 2003). Wichtig ist zu aller erst, dass der Kandidat sich wohl fühlt. Sinnvollerweise kann man wie folgt vorgehen: Man beginnt mit einer leichten offenen Aufwärmfrage (Wie geht es Ihnen? …). Geschlossene Fragen (Ja/Nein-Frage, Entscheidungsfrage) eigenen sich, um Themen zu etablieren, denn sie geben dem Prüfling Zeit, sich auf ein Thema einzustellen („Haben Sie dieses Jahr Urlaub gemacht?). Daran sollten sich offene Fragen (W-Fragen, die möglichst wenig vorgeben) oder besser explizite Aufforderungen („Erzählen Sie mir bitte von Ihrem letzten Urlaub!“) anschließen. Ein regelmäßiges verbales Feedback („aha …“) mit Interesse zeigenden offenen Anschlussfragen oder -aufforderungen („Und warum haben Sie dieses Hotel gewählt?“, „Erzählen Sie mir mehr über das Hotel.“, „Bewerten Sie die Qualität des Hotels!“) helfen dem

Kandidaten, sein Thema weiter zu entwickeln. Neue Themen seitens des Prüfers sollten möglichst an vom Prüfling angeschnittene Themen anschließen.
Zu vermeiden sind:

- Doppel-, Reihen-, Kettenfragen – zu viele Fragen auf einmal verwirren den Prüfling;
- Fortgesetzte Ergänzungsfragen lassen keinen Dialog zustande kommen
- („Nasepulfragen“);
- Schein- und Suggestivfragen („Sind Sie nicht auch der Meinung, dass“);
- Echo („Sie waren also schon einmal in Deutschland“) – dies wird oft vom Prüfling als Bitte um Bestätigung verstanden und mit „ja“ oder „nein“ beantwortet, führt eben nicht zu weiteren Ausführungen.

Reparaturen seitens des Prüfers, z.B. eine Hilfe bei einer Vokabel, wenn sie umschrieben wird oder nach ihr in der Zielsprache gefragt wird, sind authentisch für Alltagskommunikation in einer Fremdsprache und sollten daher selbstverständlich sein.
Eine **Prüfung in Paaren** ermöglicht Gesprächstypen, die im klassischen Interview nicht oder nicht authentisch möglich sind, wie z.B. Rollenspiele, Instruktion zum Zeichnen einer Grafik (und damit prüfen der Mikrofunktion Diskurssteuerung und Reparatur). Außerdem erlaubt sie der prüfenden Person, wenn sie gleichzeitig Gesprächsleiter und Prüfer ist, zeitweilig aus der Gesprächsleiterrolle zu schlüpfen und sich deshalb stärker auf die Leistungen der Prüflinge zu konzentrieren. Wichtig ist hierbei, dass die Rollen den Kandidaten unmissverständlich klar gemacht werden müssen. Außerdem muss die Struktur der Aufgabe sicherstellen, dass die Teilnehmer möglichst gleiche Chancen haben, zu Wort zu kommen – dies ist in Rollenspielen gut möglich. Hilfreich ist eine Vorselektion der Paare auf möglichst ähnlichem Niveau, um Chancengleichheit zu schaffen. Ist dies nicht möglich, ist es als Absicherung sinnvoll, auch in Paarprüfungen kurze Interaktionen der einzelnen Prüflinge mit dem Prüfer einzuplanen, um eventuelle Ungleichgewichte zwischen den Prüflingen ausgleichen zu können.

(3) Als Aufgabentypen kommen bei mündlichen Prüfungen Sprechaufgaben mit offenem Ausgang oder strukturierte Sprechaufgaben zum Einsatz (Luoma 2004: 47ff.). **Sprechaufgaben mit offenem Ausgang** geben das Thema und eine Makrofunktion vor, schränken aber den Weg und das Ergebnis des Gesprächs nicht ein. Daher lassen sie sich einteilen nach ihren Makrofunktionen:

- Beschreibung eines Gegenstandes,
- Diskussion einer These mit dem Prüfer/ Mitprüfling,
- Rollenspiel als Mischung von Diskurstypen,
- vorbereiteter Monolog.

Ein vorbereiteter Monolog ist nur sinnvoll, wenn er integriert in eine Präsentationsprüfung mit Mediennutzung (z.B. Folien) geprüft wird. Dies kann besonders für Fremdsprachen für den Beruf oder für akademische Zwecke relevant sein, sollte dann aber auch ausführlich unterrichtet werden.
Strukturierte Sprechaufgaben geben zusätzlich Verlauf, Inhalt oder Ergebnisse einer Aufgabe vor und haben daher den Vorteil der besseren Vergleichbarkeit und Reliabilität. Typisch sind diese Aufgaben unter anderem für Prüfungen die mittels Band/Computer automatisiert durchgeführt werden. Zu den strukturierten Sprechaufgaben gehört auch das Vorlesen einer Textpassage. Diese Aufgabe ist allerdings hoch problematisch, da von ihr nicht auf die Aussprache des Kandidaten im freien Gespräch geschlossen werden sollte,

schließlich ist das Vorlesen eine andere Fertigkeit und die Aussprache hier erfahrungsgemäß oft schlechter als bei freiem Sprechen. Eine weitere Möglichkeit bieten **Satzwiederholungen**. Hier hört der Kandidat Sätze und wiederholt sofort. Die Sätze werden zunehmend länger, daher liegt der Fokus dieses Tests auf der Sprachverarbeitung und der Erinnerung, weshalb diese Methode eher in psycholinguistische Versuchen Anwendung findet. Die **Vervollständigung von Satzfragmenten** oder das **Beantworten von Fragen mittels vorgegebener Fakten** prüft das Kontextverständnis und die Lexiko-grammatische Kompetenz der Kandidaten, während die Aufforderung, auf eine vorgegebene Äußerungen adäquat zu reagieren, situatives und pragmatisches Wissen prüft.

	Sprechen	
	An Gesprächen teilnehmen	**Zusammenhängendes Sprechen**
C2	Ich kann mich mühelos an allen Gesprächen und Diskussionen beteiligen und bin auch mit Redewendungen und umgangssprachlichen Wendungen gut vertraut. Ich kann fließend sprechen und auch feinere Bedeutungsnuancen genau ausdrücken. Bei Ausdrucksschwierigkeiten kann ich so reibungslos wieder ansetzen und umformulieren, dass man es kaum merkt.	Ich kann Sachverhalte klar, flüssig und im Stil der jeweiligen Situation angemessen darstellen und erörtern; ich kann meine Darstellung logisch aufbauen und es so den Zuhörern erleichtern, wichtige Punkte zu erkennen und sich diese zu merken.
C1	Ich kann mich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen. Ich kann die Sprache im gesellschaftlichen und beruflichen Leben wirksam und flexibel gebrauchen. Ich kann meine Gedanken und Meinungen präzise ausdrücken und meine eigenen Beiträge geschickt mit denen anderer verknüpfen.	Ich kann komplexe Sachverhalte ausführlich darstellen und dabei Themenpunkte miteinander verbinden, bestimmte Aspekte besonders ausführen und meinen Beitrag angemessen abschließen.
B2	Ich kann mich so spontan und fließend verständigen, dass ein normales Gespräch mit einem Muttersprachler recht gut möglich ist. Ich kann mich in vertrauten Situationen aktiv an einer Diskussion beteiligen und meine Ansichten begründen und verteidigen.	Ich kann zu vielen Themen aus meinen Interessengebieten eine klare und detaillierte Darstellung geben. Ich kann einen Standpunkt zu einer aktuellen Frage erläutern und Vor- und Nachteile verschiedener Möglichkeiten angeben.
B1	Ich kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Ich kann ohne Vorbereitung an Gesprächen über Themen teilnehmen, die mir vertraut sind, die mich persönlich interessieren oder die sich auf Themen des Alltags wie Familie, Hobbys, Arbeit, Reisen, aktuelle Ereignisse beziehen.	Ich kann in einfachen zusammenhängenden Sätzen sprechen, um Erfahrungen und Ereignisse oder meine Träume, Hoffnungen und Ziele zu beschreiben. Ich kann kurz meine Meinungen und Pläne erklären und begründen. Ich kann eine Geschichte erzählen oder die Handlung eines Buches oder Films wiedergeben und meine Reaktionen beschreiben.
A2	Ich kann mich in einfachen, routinemäßigen Situationen verständigen, in denen es um einen einfachen, direkten Austausch von Informationen und um vertraute Themen und Tätigkeiten geht. Ich kann ein sehr kurzes Kontaktgespräch führen, verstehe aber normalerweise nicht genug, um selbst das Gespräch in Gang zu halten.	Ich kann mit einer Reihe von Sätzen und mit einfachen Mitteln z. B. meine Familie, andere Leute, meine Wohnsituation meine Ausbildung und meine gegenwärtige oder letzte berufliche Tätigkeit beschreiben.

A1	Ich kann mich auf einfache Art verständigen, wenn mein Gesprächspartner bereit ist, etwas langsamer zu wiederholen oder anders zu sagen, und mir dabei hilft zu formulieren, was ich zu sagen versuche. Ich kann einfache Fragen stellen und beantworten, sofern es sich um unmittelbar notwendige Dinge und um sehr vertraute Themen handelt.	Ich kann einfache Wendungen und Sätze gebrauchen, um Leute, die ich kenne, zu beschreiben und um zu beschreiben, wo ich wohne.

Abbildung 20: GeR Deskriptoren für eine Selbstevaluation der Schreibkompetenz (Trim et al. 2009: 36)

Abbildungsverzeichnis

Bibliographie

Alderson, J. Charles (2000): *Assessing reading*. Cambridge: Cambridge University Press. (=*Cambridge language assessment series*).

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

Bolton, Sibylle (1996): *Probleme der Leistungsmessung*. *Lernfortschrittstests in der Grundstufe.* Berlin: Langenscheidt. (=*Fernstudienprojekt zur Fort- und Weiterbildung im Bereich Germanistik und Deutsch als Fremdsprache : Teilbereich Deutsch als Fremdsprache* 10).

Broadfoot, Patricia/Osborn, Marilyn/Planel, Claire (2000): *Promoting quality in learning*. *Does England have the answer?* London: Cassell.

Brown, Annie (2003): "Interviewer variation and the co-construction of speaking proficiency". *Language Testing* 1/20: 1–25.

Buck, Gary/Tatsuoka, Kumi/Kostin, Irene (1997): "The Subskills of Reading: Rule-space Analysis of a Multiple-choice Test of Second Language Reading Comprehension". *Language Learning* 3/47: 423–466.

Bygate, Martin (1987): *Speaking*. Oxford: Oxford University Press. (=*Language teaching*).

Chomsky, Noam (1969, 1965): *Aspects of the theory of syntax*. Cambridge: M.I.T. Press.

Council of Europe (ed.) (2009): *Relating Language Examinations to the Common European Framework of Reference for Languages*. *Learning, Teaching, Assessment (CEFR)*. Strasbourg: Council of Europe, Language Policy Division.

Dahlhaus, Barbara (1994): *Fertigkeit Hören*. Berlin ; München ; Leipzig ; Wien ; Zürich ; New York: Langenscheidt. (=*Fernstudienprojekt zur Fort- und Weiterbildung im Bereich Germanistik und Deutsch als Fremdsprache : Teilbereich Deutsch als Fremdsprache* 5).

Dlaska, Andrea/Krekeler, Christian (2009): *Sprachtests*. *Leistungsbeurteilungen im Fremdsprachenunterricht evaluieren und verbessern*. Baltmannsweiler: Schneider-Verl. Hohengehren.

Doyé, Peter (1988): *Typologie der Testaufgaben für den Unterricht Deutsch als Fremdsprache*. Berlin ;, New York: Langenscheidt.

Fulcher, Glenn (1999): "Assessment in English for academic purposes: putting content validity in its place". *Applied Linguistics* 2/20: 221–236.

Gnutzmann, Claus/Kiffe, Marion (1993): "Mündliche Fehler und Fehlerkorrekturen im Hochschulbereich. Zur Einstellung von studierenden der Anglistik". *Fremdsprachen lehren und lernen* 22: 91–108.

Grotjahn, Rüdiger (2000a): *Studieneinheit Leistungsmessung und Leistungsbeurteilung*. *Kapitel 1: Grundlagen*. www.uni-leipzig.de/herder/temp/lehrende/tschirner/testen/grundlag.pdf.

Grotjahn, Rüdiger (2000b): *Studieneinheit Leistungsmessung und Leistungsbeurteilung. Testen der Fertigkeit Hörverstehen.* www.uni-leipzig.de/herder/temp/lehrende/tschirner/testen/hoeren.pdf (22.11.2012).

Grotjahn, Rüdiger (2000c): *Studieneinheit Leistungsmessung und Leistungsbeurteilung. Testen der Fertigkeit Leseverstehen.* www.uni-leipzig.de/herder/temp/lehrende/tschirner/testen/Lesen.pdf.

Henrici, Gert/Herlemann, Brigitte (1986): *Mündliche Korrekturen im Fremdsprachenunterricht*. München: Goethe-Inst. (=*Materialien zur Lehrerfortbildung*).

Hughes, Arthur (2003): *Testing for language teachers*. Cambridge: Cambridge University Press. (=*Cambridge language teaching library*).

Huneke, Hans Werner (1995): "Aus Fehlern lernen? Schriftliche Fehlerkorrekturen im DaF-Unterricht auf der Mittelstufe". *Mit einem Vorschlag für den Unterricht. In: Runa. Revista Portuguesa de Estudos Germanisticos* 24/23: 467–499.

Huneke, Hans-Werner/Steinig, Wolfgang (2010): *Deutsch als Fremdsprache*. *Eine Einführung*. Berlin: Erich Schmidt.

Ingenkamp, Karlheinz/Lissmann, Urban (2008): *Lehrbuch der pädagogischen Diagnostik*. Weinheim; Basel: Beltz.

Karran, Terence (2005): "Pan-European Grading Scales: Lessons from National Systems and the ECTS". *Higher Education in Europe* 1/30: 5–22.

Kleppin, Karin (2003): *Fehler und Fehlerkorrektur*. Berlin, München, Wien, Zürich, New York: Langenscheidt.

Kleppin, Karin/Königs, Frank G. (1997): *Der Korrektur auf der Spur*. *Untersuchungen zum mündlichen Korrekturverhalten von Fremdsprachenlehrern*. Bochum: Brockmeyer. (=*Manuskripte zur Sprachlehrforschung* 34).

Luoma, Sari (2004): *Assessing speaking*. Cambridge: Cambridge University Press.

North, B./Schneider, G. (1998): "Scaling descriptors for language proficiency scales". *Language Testing* 2/15: 217–262.

Pollitt, Alastair B. Hutchinson Carolyn J. (1986): "The validity of reading comprehension tests. What makes questions difficult?". In: Vincent, Denis/Pugh, A. K./Brooks, Greg (eds.): *Assessing reading*. *Proceedings of the UKRA Colloquium on the Testing and Assessment of Reading*. London: Macmillan Education: 41–61.

Quetz, Jürgen (2003): "A1 – A2 – B1 – B2 – C1 – C2. Der Gemeinsame europäische Referenzrahmen". *Deutsch als Fremdsprache* 1/40: 42–48.

Quetz, Jürgen (2010): "Der Gemeinsame europäische Referenzrahmen als Grundlage für Sprachprüfungen. Eine kritische Beschreibung des Status quo". *Deutsch als Fremdsprache* 4/47: 195–202.

Rost, Michael (1990): *Listening in language learning*. London: Longman. (=*Applied linguistics and language study*).

(2004): *Schulgesetz für das Land Berlin*. *SchulG*.

Smith, William L. et al. (1985): "Some Effects of Varying the Structure of a Topic on College Students' Writing". *Written Communication* 1/2: 73–89.

Solmecke, Gert (2000): "Faktoren der Schwierigkeiten von Hörtests.". In: Bolton, Sibylle (ed.): *TESTDAF*. *Grundlagen für die Entwicklung eines neuen Sprachtests ; Beiträge aus einem Expertenseminar*. Köln: Gilde Verl: 57–76.

Sullivan, Kirk P. H. (2002): "Credit and Grade Transfer within the European Union's SOCRATES Programme: Unity in diversity or head in the sand?". *Assessment & Evaluation in Higher Education* 1/27: 65–74.

Trim, John et al. (2009): *Gemeinsamer europäischer Referenzrahmen für Sprachen. Lernen, lehren, beurteilen ; [Niveau A1, A2, B1, B2, C1, C2]*. Berlin: Langenscheidt.

Weigle, Sara Cushing (2002): *Assessing writing*. Cambridge: Cambridge University Press. (=*Cambridge language assessment series*).

Weir, Cyril J. (2005): "Limitations of the Common European Framework for developing comparable examinations and tests". *Language Testing* 3/22: 281–300.

Wisniewski, Katrin (2010): "Bewertervariabilität im Umgang mit GeR-Skalen. Ein- und Aussichten aus einem Sprachtestprojekt". *Deutsch als Fremdsprache* 3/47: 143–149.