Ulrike Arras
Hagen (D)

Testen und Beurteilen des Leseverstehens in der Fremdsprache

Da anni Ulrike Arras è attiva nel campo dei test per il tedesco L2 (TestDaf), un esame specifico indirizzato agli studenti stranieri che intendono iscriversi ad un'università tedesca. Muovendo da questa esperienza, l'autrice delinea alcuni importanti principi per la misurazione di competenze in lettura che dovrebbero aiutare gli insegnanti a scegliere le forme d'esame adeguate. Anzitutto ritiene importante la distinzione generale tra compiti che mirano all'esercitazione della comprensione del testo e compiti funzionali invece alla valutazione di tali capacità. Inoltre occorre tenere in considerazione l'obiettivo la valenza pedagogica di un test. Ne conseguono indicazioni per lo sviluppo dei test che concernono in particolare i criteri di scelta dei testi, la forma degli esercizi e la determinazione del livello di difficoltà. L'ampio spettro di possibilità di valutazione può contribuire alla relativizzazione dell'importanza della valutazione stessa. L'articolo conclude con una "checklist" per lo sviluppo di compiti per la valutazione della comprensione del testo, strutturata in relazione al pubblico mirato, alla scelta dei testi, alla definizione degli item e ai criteri di valutazione. (red.)

1. Vorbemerkung

Die Konzipierung von Tests zur Messung der Lesekompetenz – so wie die Konzeption jedweden Tests – erfordert einige grundlegende Überlegungen und Entscheidungen, die der folgende Beitrag skizzieren möchte. Neben dem Testkonstrukt, welches entscheidend von der Zielgruppe und dem Testziel abhängt, ist das Format der Prüfung, also die so genannten *test specifications* wie Testinhalt, Itemtyp und Beurteilungsverfahren etc. festzulegen. Im Gegensatz zu Testinstitutionen, die standardisierte Tests entwickeln, stehen Lehrkräfte vor der Frage, welche Testformen für den Kontext Unterricht (*classroom assessment*) und welche Testformen hingegen eher für formelle Tests geeignet sind. Die hier angebotenen Entscheidungshilfen möchten nicht zuletzt für die prinzipiellen Unterschiede zwischen Aufgaben sensibilisieren, die Leseverstehen trainieren und solchen, die Leseverstehen testen.

2. Ziele, Inhalte und Formate

Anders als bei der Überprüfung der Schreibkompetenz oder der mündlichen Kommunikationsfähigkeit entzieht sich der Verstehensprozess beim Hören und beim Lesen einer direkten Beobachtung: „The process is normally silent, internal, private." (Alderson 2000: 4)[1]. Dieser Befund hat entscheidende Auswirkungen auf die Testkonstruktion. Denn während wir beispielsweise bei der Messung von Schreibkompetenz Schreiben elizitieren und anhand des Produkts (eines schriftlich fixierten Textes) direkt die Leistung beobachten und evaluieren können (Performanztest), sind wir im Falle der Messung von Lesekompetenz auf indirekt erhobene Daten angewiesen (Kompetenztest). So können wir beispielsweise anhand eines Multiple-Choice-Tests zur Messung von Lesekompetenz nur eine Aussage über die Lesekompetenz treffen, wenn dem Testverfahren eine Hypothese zugrunde liegt über den Zusammenhang von sprachlicher und kommunikativer Kompetenz. Das bedeutet: Die Testaufgabe ist so zu konstruieren, dass sie das zugrunde liegende Testkonstrukt (z. B. die kommunikative Kompetenz des Verarbeitens bzw. Verstehens von bestimmten Detailinformationen eines Lesetexts) zu erfassen vermag.

Der Bestimmung von Testziel und *test specifications* dienlich ist die Heranziehung einiger Begriffe und Dichotomien. Eine gerade für Lehrkräfte zentrale Unterscheidung ist jene des formellen versus informellen Testens. Formelle Tests, insbesondere wenn es sich um so genannte *High-Stakes-Tests* handelt, also Prüfungen, von deren Ergebnis für das Individuum vieles abhängt, etwa ein Stipendium oder – um das Beispiel TestDaF anzuführen[2] – der Zugang zu einem Studium an einer deutschsprachigen Hochschule, müssen zentrale Testgütekriterien berücksichtigen. Ein hohes Maß an Standardisierung ermöglicht es, diese Gütekriterien zu erfüllen, vor allem Reliabilität und Objektivität, aber auch Praktikabilität und Ökonomie.[3] Standardisierte Tests berücksichtigen jedoch keine individuellen Faktoren (Lernbiografie, Herkunftssprache etc.), um die genannten Kriterien einzuhalten. Die Aussagekraft standardisierter Tests kann für Lehrkräfte daher durchaus begrenzt sein, denn sie messen nicht

unbedingt solche Aspekte, die für die Optimierung von Unterricht, Lernberatung etc. von Belang sind. Wenn beispielsweise ein standardisierter Test Lesen als das Verstehen einfacher Sätze operationalisiert, „then the other aspects of student's ability to reading in context in a wider sense are left out" (Leung, 2005: 871). Die unterrichtliche Praxis hingegen ermöglicht die Individualisierung von Tests etwa zur Erfassung des Lernfortschritts (*achievement tests*) einer Klasse oder zur Erstellung individueller Profile in Form von Diagnosetests, um SchülerInnen gezielt in ihrem Lernen zu unterstützen oder zu beraten. Vor diesem Hintergrund spielt *classroom embedded testing* oder *classroom assessment* eine wichtige Rolle: Es bietet neben „close-up information on student learning in context" auch Hinweise, die für die Entwicklung bzw. Überarbeitung didaktischer und curricularer Ansätze erforderlich sind. Leung kommt daher zu dem Schluss, Leistungsmessung im unterrichtlichen Kontext spiele „an important and indispensable educational role that is quite different from standardized testing" (Leung, 2005: 885). Freilich können informelle, nicht standardisierte Tests zentrale Gütekriterien nur unzureichend einhalten. Dies betrifft insbesondere die Reliabilität, denn in der Regel stehen kaum genügend Ressourcen zur Verfügung, um neben der Testkonzeption und -erstellung auch die Erprobung von Aufgaben zu ermöglichen und die Qualität der Aufgaben mittels statistischer Verfahren abzusichern. Vor diesem Hintergrund ist es nötig, die begrenzte Aussagekraft derart ermittelter Leistungen stets im Auge zu behalten; die Ergebnisse dieser im unterrichtlichen Geschehen integrierten Leistungsmessung dürfen somit nicht für „public comparison" oder für „reporting purposes" eingesetzt werden (Leung, 2005: 885).

Eine weitere für Lehrkräfte wichtige Differenzierung ist jene zwischen bezugsgruppen- und kriterienorientiertem Testen. Beim bezugsgruppenorientierten Vorgehen, zuweilen auch als normorientiertes Testen – *norm-referenced* (Hughes, 2003) – bezeichnet, wird die individuelle Leistung mit den Ergebnissen der anderen Prüflinge verglichen, so dass die erreichbaren Niveaustufen oder „Bestehensgrenzen" erst nach Durchführung der Prüfung bzw. erst anhand der Ergebnisse festgelegt werden. Beim kriterienorientierten Testen (*criterion-referenced*, Hughes, 2003) wird die individuelle Leistung hingegen in Bezug auf zuvor festgelegte Kriterien – beispielsweise Skalen oder festgesetzte Kompetenzniveaus – beurteilt. Dabei ist eine Justierung der Bestehensgrenzen oder erreichbaren Kompetenzniveaus unter Einbezug der tatsächlich gezeigten Leistungen nicht vorgesehen. Auch hier ist die Entscheidung für oder gegen das hinsichtlich der Testgütekriterien zu favorisierende kriterienorientierte Testen abhängig vom Testziel: Sollen die Ergebnisse des Tests allein im Unterricht verwendet werden, so ist auch aus lernpsychologischen und zeitökonomischen Gründen ein normorientiertes Vorgehen durchaus vertretbar. Sollen die Ergebnisse jedoch Grundlage für kurs- oder institutionsübergreifende curriculare Entscheidungen sein, so sollte das kriterienorientierte Verfahren herangezogen werden.

Das Testziel hat auch Konsequenzen für die Frage, welches Format einem Test zugrunde gelegt wird: Soll es um integriertes (*integrated*) oder um isoliertes (*discrete-point* oder *analytic*) Testen gehen? Discrete-point-Tests „zielen auf die Messung der Kenntnis spezifischer isolierter sprachlicher Phänomene" (Grotjahn, 2003: 37), d. h. das zu überprüfende Phänomen wird weitgehend eingegrenzt, etwa auf eine Teilfertigkeit oder die Beherrschung bestimmter sprachlicher Strukturen. Es handelt sich meistens um geschlossene Aufgabentypen wie Multiple-Choice- oder Zuordnungsaufgaben, deren Auswertung als ökonomisch und objektiv gilt. Bei integrativen Tests hingegen geht es darum, „to gain a much more general idea of how well students read" (Alderson, 2000: 207). Sie entsprechen eher „dem tatsächlichen Sprachgebrauch in realen Kommunikationssituationen" (Bolton 1996: 103) und sollten daher gerade im Unterricht nicht fehlen. Integrative Tests versuchen, verschiedene für das Leseverstehen relevante Handlungen zu aktivieren. Gemeinhin werden komplexe Aufgaben wie etwa die schriftliche Zusammenfassung eines Lesetextes in der Zielsprache (ein Aufgabentyp, der nicht allein Lesekompetenz, sondern auch die schriftliche Ausdrucksfähigkeit sowie strategisches Wissen erfordert und ein aufwändiges Auswertungssystem verlangt), teils aber auch C-Tests oder *cloze tests* als solche integrativen Aufgaben angesehen.[4] Die Entscheidung für isoliertes oder integriertes Testen und damit die Konzipierung der Leseverstehensaufgaben basiert auf folgenden Einzelaspekten: den zu messenden Lese- und Verstehensstrategien, der inhaltlichen bzw. thematischen und sprachlichen Gestaltung des Textes, der Textsorte, dem Itemtyp, mithilfe dessen das Verstehen beurteilt werden soll, und dem Beurteilungsverfahren. Hieraus ergeben sich Konsequenzen für die Testkonstruktion, wie im Folgenden gezeigt werden soll.

3. Testkonstruktion

3.1 Textauswahl

Selbstredend nimmt der Lesetext in einer Leseverstehensaufgabe eine zentrale Rolle ein. Die Textauswahl muss daher anhand von Kriterien erfolgen, welche bestimmt werden durch das Testkonstrukt, das Testziel und die zu überprüfenden Sprachhandlungen. Zu nennen sind insbesondere Merkmale wie die sprachliche Gestaltung des

Textes (Register, syntaktische Strukturen etc.), die Textsorte, aber auch die Textlänge usw. Ein in diesem Kontext häufig diskutiertes Anliegen ist die Authentizität. Zwar ist ein hohes Maß an Authentizität – nicht nur des Textes selbst, sondern auch der damit verbundenen Sprachhandlungen – wünschenswert, nichtsdestotrotz erweist sich die Einhaltung dieses Gütekriteriums gerade bei der Konstruktion von Testaufgaben oft als problematisch.[5] Dem Forschungsstand nach erscheint zumindest eine gemäßigte Authentizität erforderlich, die sich auf die einzuhaltenden Textmerkmale beschränkt, so dass die Texte hinsichtlich ihrer sprachlichen und textstrukturellen Merkmale authentisch „aussehen". Beispielsweise kann ein Zeitungstext angesichts der begrenzten Testzeit u. U. nicht in voller Länge verwendet werden, eine Kürzung oder Didaktisierung bzw. seine Einrichtung als Basis einer Testaufgabe ist also nötig. Dies erscheint legitim, so lange gewährleistet wird, dass die charakteristischen Merkmale des zugrunde gelegten Textes (etwa Merkmale journalistischer Texte) beibehalten werden.

3.2 Itemformate

Die Wahl der geeigneten Itemformate ist ebenfalls von Bedeutung, nicht allein im Hinblick auf das Testziel, sondern auch im Hinblick auf die Testgütekriterien. Prinzipiell können geschlossene, halboffene und offene (mehr oder weniger steuernde) Aufgabentypen unterschieden werden.[6] In standardisierten Tests wird das Leseverstehen meist mit Hilfe geschlossener oder halboffener Formate getestet, insbesondere Multiple-Choice-Aufgaben, Zuordnungsaufgaben, Aufgaben in Form von Alternativantworten oder Kurzantworten zu Fragen zum Lesetext. In nicht-standardisierten Tests, vor allem bei der Messung von Lesekompetenz im Unterricht, werden eher offene Aufgabentypen verwendet. Hierbei sind oftmals inhaltliche Fragen zum Lesetext zu beantworten; aber auch weit komplexere Aufgaben, die über das Leseverstehen hinausreichende kognitive Operationen erfordern, sind üblich wie z. B. die bereits erwähnte Zusammenfassung eines Lesetextes, also eine Textreproduktion. Ein gerade auch an der Hochschule (in philologisch geprägten Studiengängen) verbreitetes Testformat stellt immer noch die Übersetzung aus der Fremdsprache in die Erstsprache dar oder auch eine inhaltliche Zusammenfassung in der Erstsprache.[7]

Bei der Entscheidung für oder gegen ein bestimmtes Aufgaben- bzw. Itemformat befindet sich die Testkonstruktion stets in einem Dilemma, denn sie muss verschiedene, sich teils zuwiderlaufende Gütekriterien berücksichtigen. Zum einen soll ein Test ein hohes Maß an Authentizität und Validität aufweisen, d. h. die durch den Test elizitierten kognitiven Operationen sollen möglichst auch für eine reale Sprachverwendungssituation relevant sein (Bachman/Palmer, 1996: 10ff.). So zeigt ein Lesetest, bei dem die wichtigsten inhaltlichen Aussagen eines zielsprachlichen Fachtextes (in der Fremdsprache oder in der Erstsprache) zusammenzufassen sind, einen hohen Grad an Augenscheinvalidität; denn Aktivitäten wie die Verarbeitung fremdsprachlicher Fachliteratur mit Hilfe eines Exzerpts – eventuell zur weiteren Verwendung in Form eines Referats – sind Sprachhandlungen, die für den Alltag von Studierenden von Belang sind. Wenn es sich bei der Zielgruppe also um Studierende handelt, kann ein solches Testdesign in Erwägung gezogen werden. Jedoch schränkt die Praktikabilität dieses Testdesigns die Qualität des Tests ein, denn die Auswertung und die Einstufung der Leistung entpuppt sich als sehr aufwändig hinsichtlich der einzusetzenden Ressourcen wie Zeit und Personal. Auch die Reliabilität kann die Güte des Tests beeinträchtigen, denn für eine reliable Beurteilung der durch den offenen Aufgabentyp elizitierten Leistung müssen Maßstäbe und Kriterien entwickelt und ggf. skaliert werden. Zudem bleibt ungeklärt, inwiefern das Produkt, die Zusammenfassung, tatsächlich auf die zu elizitierenden Lesestrategien schließen lässt. Umgekehrt sind geschlossene Aufgabentypen wie Multiple-Choice-Tests zwar hoch reliabel und vor allem im Falle von maschineller Unterstützung bei der Auswertung der Ergebnisse auch sehr ökonomisch, jedoch kann ggf. die Validität eingeschränkt sein, weil anhand des Ergebnisses (ein Punktscore) nur bedingt auf die tatsächlich zugrunde liegende Leseverstehenskompetenz bzw. die verwendeten (und zu testenden) Lesestrategien geschlossen werden kann. Jeder Itemtyp bringt daher Vorteile, aber auch Nachteile mit sich. Der Konzipierung eines (Lese-)Tests dienlich ist daher der von Bachman und Palmer (1996: 17ff.) entwickelte Ansatz der *test usefulness*, die sich aus dem Zusammenspiel aller Testgütekriterien (Bachman und Palmer nennen hier Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Impact und Praktikabilität) ergibt.

4. Schwierigkeitsbestimmung

Insbesondere für *High-Stakes-Tests* von zentraler Bedeutung ist die Konstanthaltung der Schwierigkeit. Das heisst: Wenn ein Test zur Messung der Leseverstehenskompetenz vorgibt, das Niveau A2 abzudecken, dann müssen die dabei eingesetzten Aufgaben diesem Niveau stets entsprechen und zwar unabhängig vom Termin ihres Einsatzes. Denn wenn Tests zu unterschiedlichen Prüfungsterminen Schwierigkeitsvarianz aufweisen, dann ist nicht nur die Aussagekraft der Ergebnisse eingeschränkt, sondern für die Prüflinge erweist sich der Test auch als unfair (nicht die Lei-

stung ist ausschlaggebend, sondern das Testereignis).

In diesem Zusammenhang erhebt sich die Frage, was überhaupt eine schwierige Leseverstehensaufgabe kennzeichnet. Textverstehen ist prinzipiell in starkem Maße abhängig von der Person der Rezipientin bzw. des Rezipienten und damit von subjektiven Faktoren wie Interesse, Lesemotivation, Lesestil, aber auch vom thematischen, kulturellen und Weltwissen, von der Vertrautheit mit der Textsorte usw. Diese Subjektivität des Leseprozesses stellt die Testerstellung vor die Aufgabe, Lesetests so zu konstruieren, dass die Verstehensziele und möglichst auch die Lesestile seitens der Prüflinge weitgehend ähnlich sind. Die Schwierigkeit hängt also von verschiedenen Faktoren ab, es handelt sich um eine komplexe „Wechselbeziehung von Merkmalen des jeweiligen Lesetextes, von Eigenschaften der zugehörigen Items, von Spezifika der Instruktion sowie von personenspezifischen Merkmalen" (Grotjahn, 2000: 23f.). Auch Erkenntnisse aus der Lesbarkeitsforschung und aus Untersuchungen zur Bestimmung von Aufgabenschwierigkeiten[8] zeigen, dass sich die Schwierigkeit von Leseverstehensaufgaben nur sehr beschränkt einschätzen lässt. Was die Festlegung der Schwierigkeitsdeterminanten (und damit die Möglichkeit, die Schwierigkeit unterschiedlicher Leseverstehensaufgaben zu kontrollieren und konstant zu halten) anbelangt, so lässt sich das Problem auf zwei grundlegende Fragen reduzieren: Was ist ein schwieriger Lesetext und was sind schwierige Fragen zur Überprüfung des Leseverstehens? Nützliche Zusammenstellungen von Schwierigkeitsdeterminanten für die Konstruktion von Lese- und mit einiger Modifikation auch von Hörverstehensaufgaben finden sich bei Freedle/Kostin zum Hörverstehen (1999: 26ff.) und bei Grotjahn (2000: 47) zum Leseverstehen. Folgende Kategorien sind hierbei zu unterscheiden:

- Faktoren, die die Schwierigkeit des Textes bestimmen,
- Faktoren, die die Itemschwierigkeit determinieren und schließlich
- Faktoren, die die Text-Item-Relation bestimmen.

Erst aus dem Zusammenspiel dieser drei Aspekte lässt sich die Schwierigkeit einer Leseverstehensaufgabe näher bestimmen. So kann ein sich durch eine hohe *type-token-ratio*[9] auszeichnender Text, der zudem ein Thema behandelt, welches den Prüflingen weitgehend unbekannt ist und darüber hinaus einen hohen Abstraktionsgrad aufweist, im Hinblick auf die Gesamtschwierigkeit der Prüfung relativiert werden, wenn die Items leicht fokussierbare Informationen abtesten, deren Verständnis zudem kein Detailverstehen erfordert. Umgekehrt kann die Schwierigkeit eines Lesetests hoch sein, auch wenn der Lesetext selbst wegen seiner sprachlichen und inhaltlichen Determinanten als eher leicht einzustufen ist, die Fragen zum Text selbst jedoch komplex sind, also beispielsweise auf implizite Informationen abzielen.

Die Schwierigkeit einer Aufgabe ist jedoch nicht allein von den Text- und Itemfaktoren abhängig, sondern auch von der Beurteilung bzw. vom Beurteilungsmaßstab. Das bedeutet: Eine schwierige Aufgabe wird leichter,

Pia Di Stefano, La cartomante.

wenn die Beurteilung milde ist. Oder umgekehrt: Auch eine hinsichtlich der genannten Faktoren leichte Aufgabe kann sich als schwierig erweisen, wenn die Maßstäbe bei der Beurteilung streng sind. Das Beurteilungsverfahren ist prinzipiell von der Wahl des Itemtyps abhängig. So erfordern offene und in geringerem Ausmaß halboffene Itemtypen die Auswertung mithilfe von Beurteilungskriterien oder Musterlösungen. Die BeurteilerInnen müssen dabei Aussagen über die Leseverstehenskompetenz anhand einer meist schriftlichen Leistung treffen: Sie müssen Text rezipieren, ggf. rekonstruieren, interpretieren, mit den Vorgaben abgleichen und schließlich ein Urteil fällen bzw. die erhobene Leistung einer bestimmten Leistungsstufe zuordnen. Bei geschlossenen Itemtypen hingegen werden die korrekt gelösten Items erfasst, der ermittelte Punktscore verweist sodann auf die erreichte (und zuvor festgelegte) Kompetenzstufe.

Ein in der Entwicklung befindliches Instrument zur Kategorisierung und Schwierigkeitsbestimmung von Testaufgaben stellt das so genannte *Manual* dar. Darin enthalten sind Raster, mit deren Hilfe vor allem Testinstitutionen ihre Sprachprüfungen den Niveaustufen des Gemeinsamen europäischen Referenzrahmens für Sprachen zuordnen können. Das *Manual* wird im Kontext des vom Europarat 2003 in Auftrag gegebenen Projekts *Relating Language Examinations to the Common European Framework of Reference for Languages – CEFR* entwickelt (nähere Informationen unter www.coe.int oder www.alte.org).

5. Üben versus Testen

Üben und Testen haben prinzipiell unterschiedliche Zielsetzungen – einerseits. Andererseits jedoch wird die Ausbildung der Lesekompetenz in einer Fremdsprache gerade dadurch optimiert und individualisiert, dass Gelerntes überprüft und evaluiert wird, sei es summativ oder formativ durch Lehrkräfte oder Institutionen, sei es durch Selbstkontrolle und Instrumente der Selbstevaluation. Allerdings stehen die Interessen der Lernenden oder der Prüflinge den Interessen der Testinstitution bzw. der testenden Person oftmals entgegen. Während Prüflinge eine Prüfung bestehen und vielleicht sogar dabei gut abschneiden wollen – unabhängig von ihren tatsächlichen Kompetenzen –, liegt es im Interesse der testenden Instanz, ein möglichst zuverlässiges und valides Bild der tatsächlichen Leseverstehenskompetenz zu erhalten, entweder um Entscheidungen hinsichtlich des Prüflings zu treffen (Kurszuweisung, Zeugnis etc.) oder auch für eine Evaluation des Unterrichts oder der eigens entwickelten Lernmaterialien. Zudem spiegeln Testaufgaben meist nicht genau die Sprachhandlungen wider, die in realen Situationen vorkommen. Mit anderen Worten: Während Testaufgaben meist so konzipiert sind, dass sie bestimmte Fertigkeiten überprüfen, ist die Welt außerhalb der Testsituation nicht nach Fertigkeiten gegliedert. Diesen Einwand vermag Testen im unterrichtlichen Kontext dahingehend aufzuheben oder zu relativieren, dass sich das *classroom assessment* solcher Evaluationsinstrumente bedient, die das Lesen integriert, also kombiniert mit anderen Fertigkeiten, überprüfen. So umfasst die Handlung „ein Referat halten" mehrere relevante Sprachhandlungen, wie z. B. das Recherchieren und Lesen einschlägiger Literatur, dabei Exzerpte anfertigen, eine Vorlesung zum Thema verfolgen und Mitschriften machen, das Referat schriftlich konzipieren und schließlich mündlich vortragen. Was die Überprüfung der Kompetenzen anbelangt, so kann die Evaluation „produktorientiert" verlaufen. Das bedeutet: Die Beurteilung erfolgt allein anhand des Produkts (ein Referat wird gehalten). Eine Beurteilung der einzelnen Fertigkeiten kann jedoch kaum valide und reliabel und damit fair erfolgen: ein insbesondere das integrierte Testen betreffendes Problem. Gerade im Kursverband sind jedoch auch kleinschrittige Evaluationen möglich, indem, auch in Gruppen- oder Paararbeit denkbar, die einzelnen Arbeitsschritte bewertet werden. Insgesamt bietet Leistungsmessung im Kontext Unterricht eine breite Palette an Evaluationsformen, so dass – auch aus Sicht der Lernenden – der Charakter des Testens relativiert werden kann.

6. Statt einer Zusammenfassung: Checkliste für die Konstruktion von Leseverstehensaufgaben

Die folgende Checkliste benennt zentrale Probleme bei der Erstellung von Testaufgaben, hier in erster Linie in Hinblick auf Leseverstehensaufgaben, erhebt jedoch keinen Anspruch auf Vollständigkeit und dient lediglich der Orientierung.[10]

Zielgruppe: Wer soll geprüft werden?

- Handelt es sich um eine den prüfenden Personen bekannte Gruppe, deren Lernweg bekannt ist?
- Handelt es sich hinsichtlich Alter, kultureller Herkunft und Lernbiografie um eine homogene Gruppe?
- Welches Alter haben die Prüflinge?
- Wie groß ist die Zielgruppe?

Testkonstrukt: Was sollen die Prüflinge unter Beweis stellen?

- Welche Operationen sind erforderlich?
- Welche Lese- und Verarbeitungsstrategien sollen verwendet werden?
- Welche Textsorten sollen zugrunde gelegt werden?
- Wie viel strategisches und zielkulturelles Wissen ist erforderlich?

Textauswahl:

- Welche Merkmale (Textsorte, Textlänge, sprachliche Strukturen, rhetorische Merkmale etc.) soll der Lesetext aufweisen?

- In welchem Verhältnis steht der Textumfang zum Umfang der zu verarbeitenden Items?
- Ist das Thema den Prüflingen bekannt? Wie viel Vorwissen ist erforderlich? Wie viel strategische Kompetenz ist notwendig?
- Wie vertraut ist den Prüflingen die Textsorte? Wie kulturspezifisch ist diese Textsorte?
- Welche Lese- und Textverarbeitungsstrategien werden durch das Testformat angestrebt und welche Strategien wenden die Prüflinge tatsächlich an?
- Sind Worterklärungen notwendig? Ist es ggf. Teil des Testkonstrukts, unbekannte Wörter kontextuell zu erschließen?
- Wie authentisch müssen die Texte sein?

Itemkonstruktion:

- Welche Itemformate sind hinsichtlich der zu elizitierenden Operationen (Lesestile, Detail- vs. Globalverstehen) angemessen?
- In welcher Sprache sind die Items zu verfassen (in der Ziel- oder Erstsprache)?
- Wie komplex (sprachlich und inhaltlich) sind die Items verfasst? Sind die Items bzw. Fragen zum Text klar und verständlich? Nicht Itemverstehen ist das erklärte Testziel, sondern Leseverstehen!
- Wie sollen die Informationen des Lesetexts verarbeitet werden? Welche Sprachhandlungen bzw. Lesestrategien sollen elizitiert werden? Und welches Itemformat ist geeignet, die zu elizitierenden Handlungen zu aktivieren?
- Wie unabhängig ist die zu lösende Aufgabe vom Weltwissen der Prüflinge? Wie unabhängig ist die zu lösende Aufgabe von anderen Items? Wie können das Ergebnis verfälschende Teststrategien kontrolliert werden (zu *test-wiseness* und *lucky guessing* s. beispielsweise Weir, 2005: 94f., zu *test-taking-strategies* s. Cohen, 2000).

Beurteilung:

- Wie kann eine zuverlässige und valide Beurteilung erreicht werden?
- Wie praktikabel muss das Auswertungsverfahren sein (geschlossene, halboffene, offene Aufgabentypen)? Welche personellen und finanziellen Ressourcen stehen zur Verfügung?
- Spiegeln die Beurteilungsmaßstäbe die angestrebten Kompetenzen wider?

Anmerkungen

[1] Zum Leseprozess sowie den Verstehensmodellen beim Lesen in der Fremdsprache s. exemplarisch Ehlers (1998) sowie Alderson (2000).

[2] Beim TestDaF (Test Deutsch als Fremdsprache) handelt es sich um eine seit 2001 weltweit administrierte Prüfung für ausländische StudienbewerberInnen, die planen, ein Studium an einer deutschsprachigen Hochschule aufzunehmen. Nähere Informationen zu Testziel und Prüfungsformat s. unter www.testdaf.de.

[3] Zu den Gütekriterien eines Tests s. ausführlich Bachman/Palmer (1996).

[4] Diskussion hierzu s. Alderson (2000: 207).

[5] Zur Diskussion des Gütekriteriums Authentizität in Testaufgaben s. insbesondere Lewkowicz (2000).

[6] Überblick, Diskussion und Beispiele für Testaufgaben zur Messung des Leseverstehens s. Doyé (1988), Alderson (2000) und Hughes (2003).

[7] S. Doyé (1988: 44), kritische Diskussion s. beispielsweise Grotjahn (2003: 103f.).

[8] Überblick s. Grotjahn (2000).

[9] Die Type-Token-Ratio bezeichnet das Verhältnis zwischen der Summe verschiedener Wörter in einem Text zur Summe aller Wörter in einem Text. Je größer dieser Wert, desto komplexer ist der Text. Mit anderen Worten: Je mehr unterschiedliche Wörter in einem Text, desto schwieriger ist er zu rezipieren.

[10] Praktische Anleitungen zur Testerstellung s. vor allem Hughes (2003), auch Davidson/Lynch (2002).

Literatur

ALDERSON, J.C. (2000): *Assessing Reading*. Cambridge, Cambridge University Press.

BACHMAN, L. F. / PALMER, A. (1996): *Language Testing in Practice*. Oxford, Oxford University Press.

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

COHEN, A. D. (2000): *Exploring strategies in test-taking: fine tuning verbal reports from respondents*, in: Ekbatani, G. / Pierson, H. (eds.): *Learner-Directed Assessment in ESL*. Mahwah, N.J. / London, Lawrence Erlbaum, p. 127-150.

DAVIDSON, F. / LYNCH, B. K. (2002): *Testcraft. A Teacher's guide to Writing and Using Language Test Specifications*. New Haven / London, Yale University Press.

DOYÉ, P. (1988): *Typologie der Testaufgaben für den Unterricht Deutsch als Fremdsprache*. Berlin.

EHLERS, S. (1998): *Lesetheorie und fremdsprachliche Lesepraxis aus der Perspektive des Deutschen als Fremdsprache*. Tübingen, Narr.

FREEDLE, R. / KOSTIN, I. (1999): *Does the text matter in a multiple-choice test of comprehension? The case of the construct validity of TOEFL's mini-talks*, in: Language Testing 16/1, p. 2-32.

GROTJAHN, R. (2000): *Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TESTDAF*, in: BOLTON, S. (ed.): *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. Köln, Gilde, p. 7-55.

GROTJAHN, R. (2003): *Leistungsmessung und Leistungsbewertung. Weiterbildungs-Masterstudiengang „Deutschlandstudien". Studienschwerpunkt: Deutsche Sprache und ihre Vermittlung*. Erprobungsfassung 6/2003. FernUniversität in Hagen (unv. Manuskript).

HUGHES, A. (2003): *Testing for Language Teachers*. Second edition, Cambridge, Cambridge University Press.

LEWKOWICZ, J. A. (2000): *Authenticity in language testing: Some outstanding questions*, in: Language Testing 17-1, p. 43-64.

LEUNG, C. (2005): *Classroom Teacher Assessment of Second Language Development: Construct as Practice*, in: HINKEL, E. (ed.): *Handbook of Research in Second Language Teaching and Learning*, p. 869-888.

WEIR, C. (2005): *Language Testing and Validation: An Evidence-based Approach*. Palgrave, Macmillan.

Internetseiten

www.alte.org
www.coe.int
www.testdaf.de

Ulrike Arras

ist Referentin für Testentwicklung im 2001 gegründeten TestDaF-Institut. Sie ist für die Schulung von Testautorinnen und Testautoren sowie Beurteilerinnen und -beurteilern von Prüfungsleistungen zuständig und führt Lehrveranstaltungen und Fortbildungen zu Fragen der Prüfungsvorbereitung und Leistungsmessung durch. Außerdem berät sie andere Testinstitutionen auf dem Gebiet des Testens fremdsprachlicher und muttersprachlicher Kompetenzen.