WAS MACHT EINE AUFGABE SCHWIERIG? SCHWIERIGKEITSDETERMINANTEN IN TESTS ZUR ÜBERPRÜFUNG MÜNDLICHER KOMPETENZEN AM BEISPIEL VON TESTDAF-AUFGABEN DES PRÜFUNGSTEILS MÜNDLICHER AUSDRUCK

Ulrike Arras (TestDaF-Institut Hagen)

1 Vorbemerkung

Zur Qualitätssicherung eines Tests gehört u.a. die Konstanthaltung der Schwierigkeit. Insbesondere bei so genannten *High Stakes*-Tests, von deren Ergebnis für die Prüfungsteilnehmenden viel abhängt – die Zulassung zum Studium beispielsweise oder ein Stipendium –, gilt, dass eine Einstufung unabhängig vom Testereignis die zugrunde liegende Kann-Beschreibung repräsentieren muss. Erst konstante Schwierigkeiten über verschiedene Testereignisse hinweg erlauben es, eine Prüfung oder einzelne Prüfungsaufgaben und -leistungen den Skalen des *gemeinsamen europäischen Referenzrahmens für Sprachen* (GER) (Europarat 2001) zuzuordnen und damit ein valides Messinstrument zur Verfügung zu stellen, welches Leistungen vergleichbar und eine Prüfung somit fair macht.[1] Der Konstanthaltung der Schwierigkeit von Testaufgaben unabhängig vom Testereignis kommt damit eine zentrale Rolle zu. In diesem Zusammenhang stellt die Überprüfung mündlicher Leistungen eine besondere Herausforderung dar. Der folgende Beitrag zeigt am Beispiel des Prüfungsteils

[1] Die Veröffentlichung des GER im Jahre 2001 hat dazu geführt, dass auch Testinstitutionen anstreben, ihre Prüfungen bzw. die Kompetenzniveaus, die sie abdecken, am Referenzrahmen zu verorten. Dies betrifft insbesondere solche, die standardisierte *High Stakes*-Tests entwickeln und administrieren. Zu nennen sind hierbei vor allem in der *Association of Language Testers in Europe* (ALTE) vertretene Testinstitutionen, zu denen auch das TestDaF-Institut zählt. Sie haben sich zur Qualitätssicherung ihrer Prüfungen auf die Einhaltung bestimmter Mindeststandards in Form eines „Code of Practice" geeinigt, der u.a. vorsieht, die Anforderungen, Inhalte der Prüfung und Beurteilung der Leistungen bei verschiedenen Testereignissen vergleichbar bzw. konstant zu halten (vgl. http://www.alte.org). Die Notwendigkeit, Sprachprüfungen zu validieren und hinsichtlich ihrer Schwierigkeit zu analysieren, um sie damit transparent und vergleichbar zu machen, hat den Europarat 2003 dazu bewogen, das Projekt zur Zuordnung von Sprachprüfungen zum GER (bekannt unter der englischen Bezeichnung *Relating Language Examinations to the Common European Framework of Reference for languages - CEFR*) ins Leben zu rufen. Ziel dieses langfristigen Projekts ist es, Handreichungen für die Testerstellung zu entwickeln. Dieses so genannte *Manual* „envisages the process of linking an exam to the CEFR in three phases: specification, standardisation and empirical validation" (Lepage/North 2006) und wird damit zu einem wichtigen Instrument, mit dessen Hilfe Testinstitutionen ihre Sprachprüfungen den Niveaustufen des GER zuordnen. Auch der TestDaF wird derzeit im Rahmen dieses Projekts einer Validierung unterzogen, um eine empirisch gesicherte Zuordnung der TestDaF-Niveaustufen zum Referenzrahmen zu ermöglichen. Hierbei werden beim so genannten ‚standard setting' die Item-Schwierigkeiten der verschiedenen Prüfungsteile eingestuft, während beim so genannten ‚bench marking' die Grenzen einer Stufe festgelegt werden. Das Projekt begann im Mai 2004 und wird voraussichtlich 2007 abgeschlossen.

Mündlicher Ausdruck des „Tests Deutsch als Fremdsprache" (TestDaF), dem ein semidirektes Testformat zugrunde liegt, wie zum einen dank unterschiedlicher Aufgabenstellungen verschiedene Kompetenzen elizitiert und verschiedene Leistungsniveaus ermittelt werden, und zum anderen mit welchen Mitteln die unterschiedlichen Aufgaben hinsichtlich ihrer Schwierigkeit justiert und konstant gehalten werden.

2. Testformate zur Überprüfung mündlicher Kompetenz

Prinzipiell wird zwischen direkten, indirekten[2] und semidirekten Testformaten unterschieden[3]. Das am weitesten verbreitete Format zur Überprüfung mündlicher Kompetenzen in der Fremdsprache ist das direkte Testen. In der Regel handelt es sich um eine ‚face-to-face'-Prüfungssituation in Form von Einzel-, Paar- oder Gruppenprüfung: Lehrkräfte oder PrüferInnen stellen Fragen, die Prüflinge antworten, mitunter werden auch Rollenspiele herangezogen. Den Vorteilen dieses Formats, bei dem Leistungen direkt beobachtbar sind – Augenschein- und Konstruktvalidität sowie Akzeptanz – stehen gewichtige Nachteile gegenüber: Auf der Hand liegt zunächst die mangelnde Objektivität der Durchführung ebenso wie der Beurteilung und die eingeschränkte Beurteilungsreliabilität, insbesondere wenn die beurteilende Person mit der Person, welche die Prüfung durchführt, identisch ist. Das Problem der „interviewer variation" erweist sich hierbei als besonders kritisch, wie auch neuere Studien anhand empirischer Daten zeigen (Brown 2005). So weist Brown nach, dass „the interviewer is very much implicated in the candidate's performance and in the construction of his or her competence" (Brown 2005: 258). Das Dilemma dieser „conversational interviews" und ihrer Beurteilungsverfahren liegt darin, dass direktes Testen mündlicher Kompetenz ein hohes Maß an Validität aufweist, da die Prüflinge hierbei ihre Fähigkeit der Interaktion in (weitgehend) authentischen Situationen unter Beweis stellen können. Andererseits macht gerade dieser Vorteil eine Standardisierung unmöglich. Denn Interaktion – also Fragen, Reaktionen, Antworten, nonverbale und semiverbale Aspekte – kann nicht standardisiert werden, d.h. ein Prüfungsgespräch kann auch bei weitgehend standardisierten Stimuli und Aufgabenstellungen, wie z. B. dem vorgegebenen Thema, kaum vorhersehbar verlaufen. „Interviewer variability" zeigt sich jedoch nicht nur in unterschiedlichen „interactional styles" auf Seiten der PrüferInnen, sondern auch in unterschiedlichen Interpretationen der eigenen

---

[2] Geschlossene Itemtypen kommen aufgrund ihrer schwachen Validität selten zum Einsatz und werden daher hier vernachlässigt. Ausführungen bei Fulcher (2004: 52) zur Dichotomie „open" versus „closed task types" sowie bei Luoma (2003: 50f.).

[3] Einen Überblick über Testformate zur Überprüfung mündlicher Kompetenzen liefern Fulcher (2003: 171ff.) sowie Luoma (2004: 47ff.), wobei die beiden Autoren die Begriffe direktes bzw. indirektes Testen nicht einheitlich verwenden.

Rolle und hinsichtlich des „amount of support they produce" (Brown 2005: 269). Darüber hinaus scheint auch die Trainierbarkeit standardisierter mündlicher Prüfungen, d.h. „interaction-based tests", begrenzt zu sein, wobei Brown (2005: 259) nichtsdestotrotz „more rigorous training and monitoring of interviewers" einfordert. Zudem schränkt das Format die Bandbreite der zu messenden Kompetenzen deutlich ein. Je nach Prüfungsdauer können ein, zwei, ggf. drei Aufgabenstellungen bzw. Stimuli präsentiert und entsprechend wenige Sprechhandlungen elizitiert werden. Meist handelt es sich ohnehin um eine einzige Sprechhandlung im Rahmen eines Interviews mit einer Prüferin oder einem Prüfer – was wiederum die Bandbreite der zu messenden Kompetenzen bzw. Kompetenzniveaus begrenzt. Angesichts dieser Einschränkungen kann das direkte Testformat schließlich auch die Fairness einer Prüfung beeinträchtigen.

Insbesondere bei weltweit administrierten und standardisierten Prüfungen wie im Falle des TestDaF, werden mittlerweile semidirekte, meist CD- oder Kassetten-gestützte Verfahren verwendet.[4] Bei semidirekten Testformaten lesen oder hören die Prüflinge „the social situation where they should imagine themselves to be, and they are asked to say what they would say in the situation" (Luoma 2004: 49).[5] Zwar ersetzen semidirekte Testformate die ‚face-to-face'-Prüfungssituation, dennoch wird „reacting in situations" elizitiert (Luoma 2004: 49). Der Input (Aufgabenstellung, Sprechaufforderung etc.) wird akustisch über einen Tonträger bzw. visuell durch das Aufgabenheft präsentiert. Die verbalen Reaktionen der Prüfungsteilnehmenden werden auf Band bzw. CD gespeichert. Die Flüchtigkeit des Gesprochenen, die bei direkten Testformaten ggf. zu Problemen in der Phase der Beurteilung führen kann, ist durch die Konservierung der Leistung relativiert, denn die Leistung steht unabhängig von BeurteilerIn, Ort und Zeitpunkt für die Beurteilung zur Verfügung. Zur Illustrierung des Formats sei folgende TestDaF-Aufgabe zitiert (Modellsatz 02, Aufgabe 5, abrufbar unter http://www.testdaf.de)[7]:

---

[4] Dieses als ‚Simulated Oral Proficiency Interview' (SOPI) bezeichnetete Testverfahren wurde zuerst in den 1980er Jahren vom Center for Applied Linguistics (CAL) in Washington, USA, entwickelt (vgl. auch Kenyon 2000).

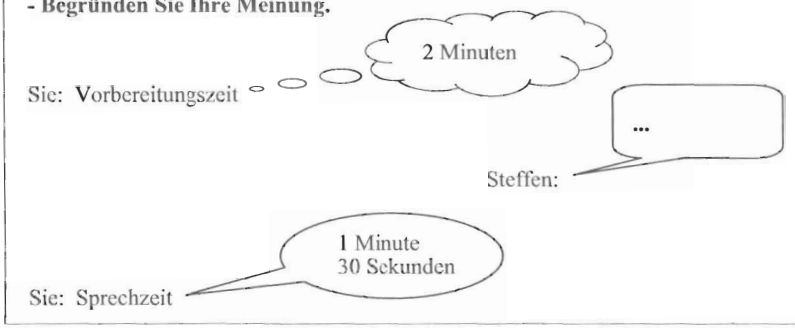
[5] Zu einer anderen Begriffsbestimmung kommt Fulcher (2003: 172ff.f.). Für ihn ist das ausschlaggebende Moment ‚the interlocutor'. Das bedeutet, ein direkter mündlicher Test liegt dann vor, wenn ein Prüfling einer Gesprächspartnerin oder einem Gesprächspartner (i.d.R. PrüferInnen oder aber andere PrüfungsteilnehmerInnen in Gruppen- oder Paarprüfungen) gegenüber verbal handelt. Indirekt nennt Fulcher somit solche Testformate, bei denen keine *face-to-face-interaction* stattfindet. Nach dieser Definition ist das kassetten- bzw. CD-gesteuerte Format beim TestDaF ein indirektes Verfahren.

[6] Es handelt sich um eine Aufgabe, die auf der TestDaF-Niveaustufe 4 angesiedelt ist.

[7] Es handelt sich um eine Aufgabe, die auf der TestDaF-Niveaustufe 4 angesiedelt ist.

Ihr Freund Steffen muss während seines Studiums ein Praktikum machen. Er hat zwei Möglichkeiten: Steffen kann das Praktikum entweder in der Firma seiner Eltern absolvieren. Oder er macht sein Praktikum in einem anderen Betrieb. Steffen fragt Sie nach Ihrer Meinung.

Sagen Sie Steffen, wozu Sie ihm raten:
- Wägen Sie die Vorteile und Nachteile der beiden Möglichkeiten ab.
- Begründen Sie Ihre Meinung.



Dieses zunächst befremdliche Testverfahren weist gewichtige Vorteile auf. Zunächst ist es das hohe Maß an Objektivität, das dieses Verfahren insbesondere für einen weltweit administrierten Test attraktiv macht. Denn nicht nur wird die Prüfung unter für alle TeilnehmerInnen weltweit gleichen Bedingungen durchgeführt (Durchführungsobjektivität), sondern auch die Beurteilung der Leistungen erfolgt ohne Ansehen der Person, ihres Aussehens, ihrer kulturellen Herkunft etc.[8], da den BeurteilerInnen allein die Tonaufnahmen zur Verfügung stehen (Beurteilungsobjektivität). Das kriterienorientierte und durch Multi-facetten-Rasch-Analysen[9] abgesicherte Beurteilungsverfahren sorgt zudem für ein hohes Maß an Reliabilität.

Freilich sind auch Nachteile des semidirekten Tests zu nennen. So wird häufig die kommunikative Angemessenheit in Frage gestellt, wenn Prüflinge „mit einer Maschine

[8] Zur Verbesserung der Objektivität ersetzt seit 2005 ein Barcode die Angabe persönlicher Daten, so dass weder Alter noch Name (und damit ggf. die kulturelle Herkunft) noch das Testzentrum aus den Prüfungsunterlagen ersichtlich sind. Damit berücksichtigt der TestDaF ein zentrales Kriterium des von der ‚Association of Language Testers in Europe' (ALTE) entwickelten ‚Code of Practice', der unter www.alte.org einsehbar ist.

[9] Da trotz Schulung und Kalibrierung Menschen unterschiedlich strenge Beurteilungsmaßstäbe anlegen, wird ein weiteres Instrument eingesetzt, um zu zuverlässigen und damit fairen Leistungsbeurteilungen zu gelangen, nämlich die Erfassung der individuellen Strenge der BeurteilerInnen mit Hilfe des Multi-Facetten-Rasch-Modells. Hierbei wird bei der Ermittlung der tatsächlich erreichten Leistungsstufe u.a. auch der Strengekoeffizient der individuellen Beurteilerin bzw. des individuellen Beurteilers einbezogen. Die endgültige Festlegung der erreichten Kompetenzstufe erfolgt unter Einbezug dieser Daten. Eine ausführliche Darstellung des Verfahrens findet sich bei Eckes 2004. Damit werden drei Aspekte bei der Ermittlung einer Leistungsbeurteilung einbezogen, nämlich dass „the score awarded to an individual on a speaking task or tasks is affected by the speaking proficiency of the individual, the difficulty of the task and the severity of the rater". (Fulcher/Márquez Reiter 2003: 322).

sprechen" sollen – eine Kritik, die insbesondere die Augenscheinvalidität und damit auch die Akzeptanz eines Tests beeinträchtigen kann. So zeigen Elder et al. (2002) anhand verschiedener Studien, dass semidirekte Testverfahren zunächst als „more difficult and/or more stressful than the live interview situation" wahrgenommen werden (Elder et al. 2002: 351). Diese Wahrnehmung scheint sich jedoch mit zunehmender Erfahrung mit dem Testformat zu relativieren, wie die Studie von Kniffka und Üstünsöz-Beurer (2001) belegt. Die Autorinnen untersuchen empirisch den Zusammenhang zwischen Vertrautheit mit dem Testformat und der mündlich gezeigten Leistung. Demnach wird das kassettengestützte Format dann problemlos angenommen, wenn sich die Prüflinge zuvor damit vertraut gemacht haben.[10] Offensichtlich gewährleistet die Vertrautheit mit dem Format, dass sich die Prüfungsteilnehmenden während der Prüfung auf die inhaltliche und sprachliche Umsetzung der Aufgabe konzentrieren und in der Folge bessere Ergebnisse erzielen.

Diese Aspekte werden jedoch durch das hohe Maß an Validität aufgefangen, denn das Format ermöglicht es, innerhalb einer Prüfung mit Hilfe verschiedener Aufgaben unterschiedliche Sprechhandlungen und damit die Verwendung einer großen Bandbreite von sprachlichen Mitteln zu elizitieren, die das Testkonstrukt widerspiegeln. Im Falle des TestDaF umfasst das Testkonstrukt solche kommunikativen Situationen, die für den Hochschulalltag relevant sind, so dass im Prüfungsteil ‚Mündlicher Ausdruck' anhand verschiedener Aufgaben unterschiedliche kommunikative Situationen präsentiert und unterschiedliche Sprechhandlungen elizitiert werden.[11] Alle Aufgabenstellungen und geforderten Sprechhandlungen haben daher einen Bezug zur Hochschule, denn das übergeordnete Prüfungsziel ist es zu überprüfen, inwieweit die KandidatInnen in der Lage sind, „verschiedene Sprechhandlungen, die im hochschulbezogenen Kontext relevant sind, angemessen zu realisieren." (http://www.testdaf.de). Die sieben Aufgaben des Prüfungsteils müssen daher unterschiedliche Sprechsituationen widerspiegeln und verschiedene Sprechhandlungen elizitieren, so dass die ermittelten Leistungen auch Niveaus zugeordnet werden. Das semidirekte Testformat bietet neben dieser differenzierten Erfassung bzw. Profilierung individueller Leistungen jedoch noch einen weiteren zentralen Vorteil: Es

[10] Auch die jeweils bei der Erprobung neuer TestDaF-Aufgaben per Fragebogen erhobenen qualitativen Daten zeigen, dass die Prüfungsteilnehmenden das kassetten- bzw. CD-gesteuerte Verfahren weit stärker akzeptieren als weitläufig vermutet wird. Zunächst wird das hohe Maß an Objektivität positiv aufgenommen. Darüber hinaus scheint die psychische Belastung im Vergleich zu den üblichen ‚face-to-face'-Testsituationen durch das semidirekte Verfahren reduziert zu werden. Insbesondere zeigt sich, dass die KandidatInnen bedingt durch das hohe Maß an Standardisierung die Möglichkeit nutzen, sich gezielt auf die Prüfung vorzubereiten, indem sie beispielsweise das Sprechen auf Band bereits vorher ausprobieren, so dass testkonstrukt-irrelevante Effekte wie Unsicherheit aufgrund der technikbasierten Prüfungsdurchführung besser kontrolliert werden können.

[11] Eine empirisch fundierte Untersuchung zu mündlichen Kommunikationssituationen und den entsprechend erforderlichen Kompetenzen vgl. Wiesmann (1999).

ermöglicht die Konstanthaltung der Schwierigkeit von Aufgaben (Stimuli, Beurteilungsmaßstäbe) unabhängig vom Testereignis, vor allem dank der Möglichkeit, sowohl die Aufgabenstellung als auch das Beurteilungsverfahren in erheblichem Maße zu standardisieren.

3. Schwierigkeitsfaktoren

Die Schwierigkeit einer Aufgabe ergibt sich aus Faktoren der Aufgabenstellung einerseits und aus Faktoren der Beurteilung andererseits. Dieses Zusammenspiel lässt sich wie folgt darstellen:

Die Leistung wird auf der Grundlage der Aufgabenstellung erbracht. Die Faktoren der Aufgabenstellung sind dabei derart, dass das zugrunde gelegte Testkonstrukt (Testziel, Kompetenzbereich etc.) widergespiegelt wird. Die Beurteilung der Leistung erfolgt anhand von Maßstäben, die dem durch die Aufgabenstellung bestimmten Prüfungsziel entsprechen müssen. Bevor die Möglichkeiten der Justierung der Schwierigkeit diskutiert werden, sollen zunächst die Schwierigkeitsfaktoren im Einzelnen skizziert werden.

3.1. Schwierigkeitsfaktoren auf Seiten der Aufgabenstellung

In standardisierten Tests sind die Schwierigkeitsdeterminanten einer Aufgabe in Form von Aufgabenmerkmalen durch das Format weitgehend festgelegt. Die Testerstellung erfolgt anhand dieser Formatvorgaben. Sie sind daher Grundlage auch der Schulung von TestautorInnen. Die Merkmalsmatrix der sieben TestDaF-Aufgaben umfasst i.W. die folgenden Schwierigkeitsdeterminanten bzw. Merkmale:[12]

- Die Aufgaben sind auf unterschiedlichen Kompetenzniveaus angesiedelt: Vorgesehen sind zwei Aufgaben auf dem Niveau TDN5, zwei auf TDN4 und drei auf TDN3. Bei der ersten Aufgabe handelt es sich um eine Warming-up-Aufgabe.

[12] Nähere Informationen zu den „task specifications" der sieben TestDaF-Aufgaben zum Mündlichen Ausdruck (vgl. http://www.testdaf.de).

- Der Input wird sprachlich (auditiv über Band sowie visuell per Aufgabenheft) präsentiert. Zwei Aufgaben stützen sich zudem auf Grafiken, die statistische Daten präsentieren.
- Von zentraler Bedeutung für die Schwierigkeit einer Aufgabe ist das Thema, über das gesprochen werden soll.
- In unmittelbarer Nähe zur Frage des Themas bzw. des Themenbereichs sind die geforderten Sprechhandlungen zu sehen. Es sind sowohl argumentative und diskursive Sprechhandlungen vorgesehen, etwa Abwägen, Begründen und Stellung beziehen oder einen Rat geben, als auch das Beschreiben eigenkultureller Phänomene sowie das Beschreiben statistischer Daten, die in einer Grafik präsentiert werden.
- Der situative Kontext, in dem sprachlich reagiert werden soll. Da es sich um Aufgaben im Hochschulkontext handelt, unterscheiden wir zwischen jenen Aufgaben, die beispielsweise in einem (Fach-)Seminar angesiedelt sind und jenen, die in einer informellen Situation, etwa im Studierendenwohnheim „spielen". Die Prüfungsteilnehmenden spielen jeweils sich selbst in einer Rolle als Student oder Studentin an einer Hochschule in Deutschland.
- Mit der Frage der Situation eng verbunden ist jene nach dem Adressatenkreis. Je nach Aufgabe spricht ein Prüfling zu einer Person oder aber zu mehreren.
- In unmittelbaren Zusammenhang hiermit ist schließlich auch das erforderliche Register zu sehen. So erfordert das Sprechen beispielsweise im Rahmen eines Kurzvortrags ein eher formelles, das Sprechen mit KommilitonInnen in informellen Situationen, wie z. B. in der Mensa, ein eher informelles Register.
- Die zeitliche Ausstattung, d.h. eine Festlegung darüber, wie viel Zeit für die Vorbereitung und wie viel für die verbale Umsetzung der Aufgabe eingeräumt wird (Vorbereitungszeit, Sprechzeit).

Diese für den TestDaF gültigen Schwierigkeitsfaktoren, können u.U. auch für andere Testformate von Bedeutung sein. Freilich sind weitere Merkmale und Systematisierungen möglich. Hilfreich sind hierbei auch jene Überlegungen von Fulcher (2003) bzw. Fulcher/Márquez Reiter (2003), die sich bei der Zusammenstellung jener psycholinguistischen Faktoren, die die Aufgabenschwierigkeit beeinflussen, auf Skehan (1998) beziehen:

- Familiar information: The more familiar the information on which a task is based, the more fluent the performance will be.

- Structured tasks: Where the task is based on a clear sequential structure there will be significantly greater fluency and accuracy.
- Complex and numerous operations: The greater the number of online operations and transformation of material that are needed, the more difficult the task. This may impact upon greater complexity, but at the expense of accuracy and fluency.
- Complexity of knowledge base: the more open the knowledge base on which a task draws, the more complex will be the language produced.
- Differentiated outcomes: as a task outcome requires more differentiated justification, the complexity of the language will increase.

(Fulcher 2003: 63)

Allerdings dienen diese Kategorien der Informationsverarbeitung nicht in erster Linie der Analyse von Testaufgaben, sondern von mündlichen „classroom activities", so dass weitere Aspekte hinzukommen, die speziell die Testsituation berücksichtigen. In seinem Forschungsüberblick nennt Fulcher (2003: 64) die folgenden Faktoren:

- „Perspective", also die Erzählperspektive, d.h. wird eine Geschichte aus der eigenen Sicht oder aber aus der Sicht einer anderen Person erzählt.
- „Immediacy" und „Adequacy", das bedeutet, ob anhand einer Vorlage, beispielsweise ein Bild, oder ohne eine solche erzählt wird, ob die Bildergeschichte, die zu versprachlichen ist, vollständig ist, oder ob fehlende Teile ergänzt werden müssen.
- Schließlich der für Prüfungen besonders wichtige Aspekt der „planning time", also die Frage, danach, ob Zeit zur Vorbereitung der Äußerung eingeräumt wird.

Insgesamt betrachtet kann festgehalten werden, dass sich die Schwierigkeit einer Aufgabe aus der Kombination verschiedener Faktoren ergibt. Das bedeutet z. B., dass wir nicht als Regel festlegen können: Je mehr Zeit zur Umsetzung der Aufgabe zur Verfügung steht, desto einfacher ist die Aufgabe. Denn ein großzügig bemessener Zeitrahmen geht einher mit einem beispielsweise thematisch komplexen Thema. Die Schwierigkeit einer Aufgabe kann somit nur im Zusammenspiel aller Faktoren justiert werden. Dies ist bei der Konzipierung, Erstellung, Durchführung und schließlich auch bei der Beurteilung (s. u.) von mündlichen Leistungen zu berücksichtigen. Die folgende Merkmalsmatrix kann daher der Orientierung bei der Justierung von Aufgabenschwierigkeiten dienen, nicht nur bei semi- sondern auch bei direkten Testverfahren wie Interview, Prüfungsgespräch etc.

Faktor	Schwierigkeit eher niedrig	Schwierigkeit eher hoch
	Situation / situative Einbettung	
Register	informell	formell
soziales Verhältnis	horizontal	vertikal / hierarchisch
sozialer Status der angesprochenen Person	niedrig, ebenbürtig	höher
sozialer Rahmen	informeller Kreis, kleine Gruppe	formelle Situation, große Gruppe
Vertrautheit mit der angesprochenen Person	groß, enge Beziehung	gering, vorgesetzt, abhängig
Rede	dialogisch	monologisch
Medium	direktes Gespräch	Telefonat
Rolle	Identifikation möglich	Identifikation nicht möglich / erschwert
Rolle	sich selbst „spielen"	eine andere Person spielen
Erzählperspektive (beim Erzählen einer Geschichte oder eines Erlebnisses)	aus der eigenen Perspektive erzählen	aus der Perspektive einer dritten Person erzählen
Antizipierter Wissensstand des Auditoriums	niedrig	hoch
Authentizität der Situation	Situation relevant / authentisch	Situation wenig relevant / authentisch
Ort	vertraut	unbekannt
	Input	
visueller Input	unterstützend (z. B. Arbeitsblatt)	inhaltsorientiert (z. B. Daten einer Grafik)
visueller Input: Sprache	vorhanden und unterstützend / erklärend	nicht vorhanden / vorhanden, aber komplex
visueller Input: Informationsdichte	niedrig	hoch
visueller Input: Textsorte	bekannt, vertraut	unbekannt
Arbeitsanweisung / situative Einbettung	klar, redundant	komplex, nicht redundant
Arbeitsanweisung / situative Einbettung: Informationsdichte	niedrig	hoch
Arbeitsanweisung / situative Einbettung: Sprache	Muttersprache	Fremdsprache
auditiver Input (z. B. Redeaufforderung)	informell	formell, distanzierend

	Thema / Problem	
erforderliches Wissen	bekannt, Allgemeinwissen, Vorwissen vorhanden	unbekannt, Spezialwissen, kein Vorwissen vorhanden
Thema / Problem – Interessantheitsgrad	hoch	niedrig
Thema / Problem	eigenkulturell	fremdkulturell
Thema / Problem	konkret	abstrakt
Themenbereich	Alltagserfahrung, persönliche Erfahrung	akademisch, beruflich, technisch, fachlich
Kreativität	nicht erforderlich	erforderlich
	Output	
Sprechhandlung	informieren, bitten, sich erkundigen, eine Geschichte erzählen	argumentieren, versachlichen, abwägen, begründen
Sprechhandlung	Wissen wiedergeben	Wissen verarbeiten
Mitteilung	persönlich	global, entpersonalisiert
Register	informell	formell
erforderliche Syntax	einfach	komplex (z. B. wissenschaftssprachliche Strukturen)
erforderlicher Wortschatz	einfach, Alltagsbereich	präzise, breit, differenziert, evtl. Terminologie
erforderliche strategische Kompetenz	gering	stark
	Zeit	
Denkzeit / Vorbereitungszeit	lang	kurz
Redezeit	lang	kurz
Redezeit / Denkzeit	selbstbestimmt	fremdbestimmt

Die Tabelle erhebt keinen Anspruch auf Vollständigkeit und kann entsprechend erweitert werden. Bei der Zusammenstellung von Schwierigkeitsfaktoren zur Konzipierung von Aufgaben, die ein bestimmtes Testkonstrukt widerspiegeln sollen, ist v.a. zu berücksichtigen, dass einzelne Faktoren Auswirkungen auf andere haben. Die Veränderung eines Aspekts zu „eher schwierig" oder „eher leicht" interagiert also mit anderen Aspekten, so dass stets die Aufgabe als Ganzes, also auch die Beurteilungsmaßstäbe und das zugrunde gelegte Testkonstrukt einzubeziehen ist. Daher zeigt die Übersicht auch, dass bestimmte Faktoren die Schwierigkeit sowohl absenken als auch erhöhen können, je nach der

persönlichen Wahrnehmung des Prüflings, aber auch je nachdem, welche Kombination sie mit anderen Determinanten einnehmen. Bei der Analyse des oben zitierten Aufgabenbeispiels können anhand der Vorgaben verschiedene Schwierigkeitsfaktoren differenziert werden, so insbesondere die folgenden Aspekte:

- Die situative Einbettung legt die *soziale Beziehung* fest, in der sich der Prüfling mit dem Gesprächspartner Steffen befindet. Beide sind Studierende, so dass es sich um ein informelles Gespräch handelt, welches entsprechend ein informelles *Register* erforderlich macht.
- Die situative Einbettung gibt das *Thema* vor. Es handelt sich zwar um ein Thema aus der studentischen Erlebniswelt, allerdings ist es auf einer Abstraktionsskala höher angesiedelt als beispielsweise das Erzählen eines Alltagserlebnisses. Das Thema sollte also *sachlich* differenziert dargestellt werden.
- Die fett gedruckte Arbeitsanweisung benennt die geforderten *Sprechhandlungen*. Einem anderen Menschen einen Rat zu geben, erfordert es, sich in die Lage der anzusprechenden Person zu versetzen und aus deren Perspektive beide Möglichkeiten diskursiv abzuwägen.

Auch die Sprechhandlung „einen Vortrag halten", um ein Beispiel zu nennen, welches in mündlichen Prüfungen gerade auch im Hochschulkontext von Relevanz ist, kann sowohl eher leicht als auch eher schwierig sein bzw. als schwierig resp. leicht empfunden werden, je nach den Koordinaten der Aufgabe. Beispielsweise bestimmen Fragen wie die folgenden die Aufgabenschwierigkeit:

- Erfordert das Thema des Vortrags spezielles Wissen oder reicht beispielsweise eigenkulturelles Wissen aus?
- Gibt es Vorgaben, die für die Äußerung verarbeitet werden müssen? Wie komplex und wie unterstützend sind sie? Schränken die Vorgaben die Äußerung ein? Oder sind sie eher fakultativ heranzuziehen?
- Wie groß ist das Auditorium? Wie viel Vorwissen haben die Zuhörenden? Handelt es sich um vertraute Personen oder handelt es sich um ein anonymes Publikum?
- Wie viel Zeit wird zur Vorbereitung eingeräumt? Wie viel Zeit bleibt dem Prüfling zur Strukturierung des Vortrags?
- Können Medien eingesetzt werden? Müssen sie eingesetzt werden?

Prinzipiell ergibt sich die Schwierigkeit einer Aufgabe aus der Aufgabe selbst sowie aus der Beurteilung der anhand dieser Aufgabe erbrachten Leistung. Die Schwierigkeit einer Aufgabe lässt sich somit nicht allein aus

der Aufgabenstellung selbst bestimmen, sondern ist stets in Verbindung mit den Beurteilungskriterien bzw. deren Interpretation zu sehen. Das bedeutet, die Schwierigkeit einer Aufgabe ist nicht absolut, denn ausschlaggebend ist nicht zuletzt der Maßstab mit dem die Leistung, die anhand der Aufgabenstellung elizitiert wurde, beurteilt wird. M.a.W.: Eine leichte Aufgabe wird zu einer schwierigen, wenn der Beurteilungsmaßstab eine entsprechende Strenge vorsieht. Oder umgekehrt: Eine hinsichtlich der Schwierigkeitsdeterminanten komplexe Aufgabe wird weniger anspruchsvoll und deckt somit ein niedrigeres Kompetenzniveau ab, wenn die Beurteilung milde ausfällt. Die schließlich durch das Urteil ausgewiesene Kompetenzstufe kann somit u. U. nicht allein an der Aufgabenstellung orientiert sein. Im Folgenden soll daher die Justierung der Schwierigkeit einer Aufgabe durch Beurteilungsmaßstab und Kalibrierung beschrieben werden.

3.2. Schwierigkeitsfaktoren auf Seiten der Beurteilung

Voraussetzung für eine standardisierte Prüfung ist ein kriterienorientiertes Beurteilungs-verfahren. Ausschlaggebend bei der Bewertung von Leistungen ist demnach nicht die (durchschnittliche) Leistung der Gesamtgruppe, d.h. alle Teilnehmenden an einem Testereignis, sondern die individuelle Leistung in Bezug auf die Anforderungen, wie sie sich in Form von skalierten Deskriptoren in den TestDaF-Niveaustufen widerspiegeln. Die Kriterien erfassen während oder nach einem ersten Hören der individuellen Leistung die Gesamtwirkung bei der Rezeption eines Textes. Es handelt sich um eine eher holistische Erfassung der Leistung. Hier werden mittels zweier Einzelaspekte die inhaltliche und phonetische Nachvollziehbarkeit der Äußerung erfasst, ohne detailliert die sprachliche und inhaltliche Umsetzung zu analysieren. Dieser holistische Zugriff erfasst somit die allgemeine kommunikative Qualität der Leistung. Danach, d.h. in der Regel nach einem zweiten Hören der Leistung, werden die sprachliche und die inhaltliche Umsetzung der jeweiligen Aufgaben beurteilt, was ein eher analytisches Vorgehen bei der Bewertung erfordert. Diese beiden Bereiche werden in Hinblick auf jeweils drei Einzelaspekte beurteilt. Damit wird das Profil einer individuellen Leistung durch insgesamt acht Einzelaspekte erfasst, wobei alle Einzelaspekte gleiches Gewicht haben.[13]

Da die Aufgaben drei unterschiedliche Leistungsniveaus abdecken, sind drei Beurteilungs-raster erforderlich, ein Raster, welches die Aufgaben bzw. Leistungen des Niveaus TDN3 erfasst, ein Raster, das zur Beurteilung jener Leistungen herangezogen wird, die durch Aufgaben auf TDN4 elizitiert wurden, und schließlich ein entsprechendes Raster auf TDN5 (Arras, erscheint). Da die Beurteilungskriterien testsatzübergreifend verfasst sind, also für alle Aufgaben, unabhängig vom Testereignis gültig sind, die Schwierigkeit einer einzelnen Aufgabenstellung jedoch auch bei starker Standardisierung nicht konstant gehalten werden kann, bedarf es weiterer Beurteilungsinstrumente: Die so genannten testsatzspezifischen Kalibrierungsunterlagen bestehen zum einen aus testsatzspezifischen

[13] Die Beurteilungskriterien im Einzelnen werden in paraphrasierter Form unter http://www.testdaf.de dargestellt.

Erläuterungen zu den einzelnen Aufgaben und zum anderen aus testsatzspezifischen beispielhaften Beurteilungen. Diese Unterlagen erhalten alle BeurteilerInnen, welche die Bewertung mündlicher Leistungen aus einem Testereignis vornehmen. Die testsatzspezifischen Erläuterungen justieren die Aufgaben, indem die Anforderungen an die jeweilige Aufgabe festgehalten und durch Beispiele illustriert werden. Zu jeder Aufgabe wird anhand von Texten aus der Erprobungsphase der Aufgaben eruiert, welche Leistungen bzw. Anforderungen die Aufgabe elizitiert und welche Maßstäbe bei der Umsetzung der jeweiligen Aufgabe anzulegen sind. Die Konstanthaltung der Aufgabenschwierigkeit über verschiedene Testereignisse und Aufgaben hinweg erfolgt also, indem festgelegt wird, inwieweit bei eher komplexen Aufgaben geringere Anforderungen an die Umsetzung zu stellen sind als bei Aufgaben, die z. B. eher einfach zu erfassende Darstellungen statistischer Daten aufweisen. Zum anderen werden ebenfalls anhand von Leistungen aus der Erprobungsphase der jeweiligen Aufgabe Texte auf unterschiedlichen Leistungsniveaus ausgewählt und durch ein ExpertInnen-Gremium, bestehend aus geschulten BeurteilerInnen und TestentwicklerInnen, bewertet. Die Urteile und entsprechende begründete Einstufungen werden schriftlich fixiert und stellen eine weitere Orientierungshilfe für alle BeurteilerInnen dar, die an der Bewertung schriftlicher bzw. mündlicher Leistungen aus einem Testereignis eingesetzt werden. Diese Kalibrierungsmaßnahmen haben zum Ziel, die Beurteilungsmaßstäbe zu bestimmen. Die konstante Interpretation dieser Maßstäbe ist zum einen wichtig, um die Reliabilität der Beurteilungen zu erhöhen, und zum anderen, um die Schwierigkeit der Aufgabe zu justieren. Die Konstanthaltung der Schwierigkeit schließlich ist erforderlich, um das Gütekriterium der Validität zu erfüllen.

4. Schlussbemerkung

Das hier skizzierte Zusammenspiel zwischen den Schwierigkeitsdeterminanten der Aufgabenstellung auf der einen und den Beurteilungsmaßstäben auf der anderen Seite ermöglicht es zum einen, testsatzspezifisch die Schwierigkeit einer Aufgabe zu justieren, und zum anderen Sorge dafür zu tragen, dass Testaufgaben bzw. Prüfungsteile unabhängig vom Testereignis weitgehend gleich schwierig – und damit fair – sind. Zusammenfassend kann somit festgehalten werden, dass für die Qualitätssicherung eines Tests – nicht nur zur Erfassung mündlicher Kompetenzen – die folgenden Aspekte von Bedeutung sind:

- Festlegung eines stets einzuhaltenden Aufgabenformats
- Gestufte Aufgabenschwierigkeit durch Festlegung der einer Aufgabe zugrunde gelegten Schwierigkeitsdeterminanten

- Konstanthaltung der Aufgabenschwierigkeit
- Kriterienorientierte Beurteilung und Operationalisierung der Beurteilungskriterien
- Justierung der testsatzspezifischen Aufgabenschwierigkeit mittels Kalibrierungsbeispielen

Wenn die genannten Kriterien erfüllt sind, dann kann nicht nur davon ausgegangen werden, dass – in Anlehnung an das geflügelte Wort von Charles Alderson – dort B2 drin ist, wo B2 drauf steht, sondern auch, dass *immer* dort B2 drin ist, wo B2 drauf steht – und damit einem zentralen Anliegen allen Testens Rechnung getragen wird, nämlich der Fairness.

5. Literaturverzeichnis

Arras, Ulrike (erscheint 2006): *Zur Revision des Subtests ‚Mündlicher Ausdruck' der Prüfung ‚Test Deutsch als Fremdsprache' (TestDaF);* in: *Studienkolleg* 10.

Brown, Annie (2005): *Interviewer Variability in Oral Proficiency Interview;* Peter Lang, Frankfurt et al..

Eckes, Thomas (2004): *Facetten des Sprachtestens. Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen;* in: Wolff, Armin et al.: *Integration durch Sprache;* Regensburg, 485-518.

Elder, C.; Iwashita, N.; McNamara, T. (2002): *Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?;* in: *Language Testing*, Vol. 19, No. 4, 347-368.

Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen;* Langenscheidt, Berlin et al..

Fulcher, Glenn (2003): *Testing Second Language Speaking;* Pearson/Longman, London et al.

Fulcher, Glenn; Márquez Reiter, Rosina (2003): *Task difficulty in speaking tests;* in: *Language Testing*, Vol. 20, No. 3, 321-344.

Kenyon, Dorry (2000): *Tape-mediated Oral Proficiency Testing: Considerations in Developing Simulated Oral Proficiency Interviews (SOPIs);* in: Bolton, Sibylle (ed.): *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar;* Goethe-Institut, München, 87-106.

Kniffka, Gabriele; Üstünsöz-Beurer, Dörthe (2001): *TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassettengesteuerten Testformats;* in: *Fremdsprachen Lehren und Lernen* 30, 127-149.

Lepage, Sylvie; North, Brian (2006): *Relating oral productions to the Common European Framework of Reference for Languages;* in: *ALTE News*, Winter 2005 / Spring 2006. www.alte.org.

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

Skehan (1998): *A Cognitive Approach to Language Learning;* Oxford University Press, Oxford.

Wiesmann, Bettina (1999): *Mündliche Kommunikation im Studium. Diskursanalysen von Lehrveranstaltungen und Konzeptionalisierung der Sprachqualifizierung ausländischer Studienbewerber*, München.

Internetseiten

Association of Language Testers in Europe: http://www.alte.org

TestDaF-Institut und -Test: http://www.testdaf.de